VYSOKÁ ŠKOLA BÁŇSKÁ – TECHNICKÁ UNIVERZITA OSTRAVA
EKONOMICKÁ FAKULTA

KATEDRA SYSTÉMOVÉHO INŽENÝRSTVÍ

Návrh a evaluace řešení aplikace asociačních pravidel při determinaci vztahů mezi produktovými skupinami

Design and Verification of Association Rules Application Solution for Determination of Relationships Between Product Groups

Student: Adam Strakoš

Vedoucí diplomové práce: Ing. Radek Němec, Ph.D.

Ostrava 2015

VŠB - Technická univerzita Ostrava
Ekonomická fakulta
Katedra systémového inženýrství

Zadání diplomové práce

Student: **Bc. Adam Strakoš**

Studijní program: N6209 Systémové inženýrství a informatika

Studijní obor: 6209T025 Systémové inženýrství a informatika

Téma: Návrh a evaluace řešení aplikace asociačních pravidel při determinaci vztahů mezi produktovými skupinami
Design and Verification of Association Rules Application Solution for Determination of Relationships Between Product Groups

Zásady pro vypracování:

1. Úvod
2. Teoretická a metodologická východiska aplikace asociačních pravidel
3. Návrh řešení pro efektivní tvorbu a aplikaci asociačních pravidel
4. Evaluace navrhovaného řešení na anonymizovaných datech
5. Závěr

Seznam použité literatury
Seznam zkratek
Prohlášení o využití výsledků diplomové práce
Seznam příloh
Přílohy

Seznam doporučené odborné literatury:

HAN, Jiawei a Micheline KAMBER. *Data Mining: Concepts and Techniques.* San Francisco: Elsevier, 2006. 743s. ISBN: 978-1-55860-901-3.
VERCELLIS, Carlo. *Business Intelligence – Data Mining and Optimization for Decision Making*. Indianapolis: John Wiley & Sons, 2009. 417 s. ISBN: 978-0-470-51138-1.
LACKO, Luboslav. *Databáze: datové sklady, OLAP a dolování dat s příklady v Microsoft SQL Serveru a Oracle*. 1. vyd. Brno: Computer Press, 2003. 486 s. ISBN 80-722-6969-0.

Formální náležitosti a rozsah diplomové práce stanoví pokyny pro vypracování zveřejněné na webových stránkách fakulty.

Vedoucí diplomové práce: **Ing. Radek Němec, Ph.D.**

Datum zadání: 21. listopadu 2014
Datum odevzdání: 25. dubna 2015

doc. Ing. Jana Hančlová, CSc.	prof. Dr. Ing. Dana Dluhošová
vedoucí katedry	*děkanka fakulty*

Prohlášení

Místopřísežně prohlašuji, že jsem celou tuto diplomovou práci včetně všech příloh vypracoval samostatně.

V Ostravě dne: ..... Podpis:

Obsah

1 Úvod

Od počátku 21. století žijeme ve světě s rapidně se zvyšujícími nároky na zpracování dat. Dnešní uspěchaná doba vyžaduje okamžité reakce na všemožné výkyvy ve finančních a jiných ukazatelích. Bez výpočetní techniky ke zpracování dat se dnes obejde málokterý podnikatelský subjekt. U maloobchodníku se smíšeným zbožím to platí dvojnásob. Chce-li maloobchodník nejen obstát, ale i růst ve svém podnikání, musí být připraven pružně reagovat na situace, které s sebou trh přináší. Dobrý a konkurenceschopný maloobchodník svá prodejní data nejen uchovává v databázových systémech, ale zná a využívá i způsoby, jak ze svých dat vytěžit co nejhodnotnější informace, které lze přeměnit ve znalosti. Tyto znalosti pak ulehčují budoucí rozhodování o obchodních a marketingových strategiích na různých úrovních řízení.

Pro zjednodušení přeměny dat na informace a znalosti se budují rozsáhlé BI systémy, v nichž jsou data uspořádána ve formě, která je příhodnější pro provádění analýz. Z dat, která jsou uložena v BI systémech, lze získat informace i jinými komplikovanějšími způsoby než je pouhé využití reportovacích nástrojů se základními funkcemi. Mezi takovéto způsoby patří právě metodologie dolování dat, která zahrnuje různé techniky, pomáhající odhalit netriviální skryté souvislosti v datech. Jejich znalost pak může být rozhodující v nelítosném konkurenčím prostředí.

Zřídkakdy je jednoduché odhalit informačně přínosné souvislosti prostřednictvím některé z technik. Tato diplomová práce je věnována právě specifickému problému v technice dolování asociačních pravidel a jejímu vztahu se sférou maloobchodu.

Cílem práce je návrh pro zlepšení procesu dolování jednodimenzionálních asociačních pravidel na transakčních datech s hierarchickými strukturami produktů z maloobchodních sítí, které povede k mj. uživatelsky snadnější identifikaci významných pravidel.

V tomto cíli je také zahrnuta aplikace části navrhovaného řešení na anonymizovaných datech s důrazem na vhodnou interpretaci nalezených pravidel pro manažery. Práce je tedy zaměřena na konkrétní případ z praxe, který s sebou přináší několik přímých požadavků, ze kterých byl cíl práce odvozen.

2 Teoretická a metodologická východiska aplikace asociačních pravidel

Praxi předchází znalost teoretických poznatků, které budou vymezeny v této kapitole. Kromě hruběji nastíněných obecnějších pojmů budou detailně vysvětleny pojmy zásadní, nepostradatelné v praktické části práce, jejichž podmnožina bude také podpořena nezbytným matematickým aparátem.

V této kapitole bude nejprve popsána disciplína zvaná dolování dat, která zahrnuje několik konkrétních metod, mezi něž se řadí metoda hledání asociačních pravidel. Dále bude podrobněji vysvětlena problematika asociačních pravidel a jejich dolování. Poslední část bude zaměřena na vztah procesů řízení maloobchodu k hledání asociačních pravidel v souvislosti s analýzou nákupního košíku.

2.1 Metoda hledání asociačních pravidel

Metoda hledání (dolování) asociačních pravidel patří mezi používané metody dolování dat. Jejím cílem je nalezení opakujících se vzorů v datech a specifikovat pravidla (asociace) v jejich vzájemných vztazích s využitím logiky a teorie pravděpodobnosti. Základy této metodiky položili již před více než dvaceti lety Agrawal a kol. (1993) v práci s názvem "Mining Association Rules between Sets of Items in Large databases". Z této práce je patrno, že byla zaměřena pouze na tzv. analýzu nákupního košíku, která měla využití u transakčních dat z maloobchodních sítí. Z pohledu na technologický vývoj až do současnosti je zřejmé, že se mezitím objevily nové oblasti, ve kterých mohly být vyvinuty také nové algoritmy pro aplikaci asociačních pravidel.

V této podkapitole bude nejprve ozřejmen pojem data mining, jakožto obor informatiky, jehož je metoda hledání asociačních pravidel součástí. Dále budou stručně popsány oblasti možné aplikace asociačních pravidel a objasněny způsoby rozdělení asociačních pravidel. Následně budou asociační pravidla v jejich základní podobě formálně definována a bude popsán v současnosti nejpoužívanější algoritmus pro jejich sestavování.

2.1.1 Data mining

Širší veřejností je na pojem „data mining“ (překl.: dolování dat) nahlíženo, jako na způsob získávání znalostí z obrovského množství dat. Han a kol. (2012) uvádějí, že z tohoto důvodu poměrně často dochází k zaměňování s pojmem „Knowledge Discovery from Data“ (KDD, překl.: získávání znalostí z dat). Pojem KDD však na druhé straně bývá popisován jako proces s iterativní posloupností kroků, jehož je dolování dat pouze součástí – což je názor většiny odborné literatury, kde však názorová jednotnost také není úplná. Pro ilustraci odlišnosti chápání tohoto pojmu jsou níže uvedeny definice dolování dat dle různých autorů:

Dolování dat je definováno jako proces hledání opakujicích se vzorů v datech (Witten a kol., 2011, s. 5).

Dolování dat v jádru zahrnuje algoritmy, které umožňují získat pohled do podstaty věci a znalosti z obrovského množství dat (Zaki a kol., 2014, s. 25).

Pojem dolování dat je proces skládající se ze sběru dat, analýzy, vývojem induktivních matematických modelů, přijetí praktických rozhodnutí a následných opatření založených na získaných znalostech (Vercellis, 2009, s. 78).

Han a kol. (2012) uvádí tyto kroky KDD:

1. Čištění dat
2. Integrace dat
3. Výběr dat
4. Transformace dat
5. Dolování dat
6. Vyhodnocení opakujících se vzorů
7. Prezentace znalostí

První čtyři kroky jsou různé formy předzpracování dat pro následné dolování. Čištění dat znamená odstranění šumu a nekonzistencí v datech, způsobené např. lidskou stránkou zasahující do dat. Integrací dat je myšlena možná kombinace dat z různých zdrojů (jejímž výsledkem může být např. vytvoření datového skladu). Výběrem dat se pak získají z databáze data, relevantní pro určitou analytickou úlohu. Následně provedením např. agregačních operací se v kroku transformace dat data konsolidují do podoby vhodné k dolování. V pátém

nejdůležitějším kroku přichází na řadu samotné dolování, kdy jsou aplikovány metody pro extrakci opakujících se vzorů, závislostí, nebo pouhých popisných charakteristik. Dále je nutno vyhodnotit získané výsledky a na základě metrik významnosti identifikovat ty, které mohou být transformovány na znalosti. V posledním kroku jsou tyto znalosti prezentovány uživateli, přičemž je kladen důraz na přívětivou vizuální formu.

V současnosti bývá v praxi i ve vědeckých kruzích více popularizován pojem dolování dat, jako celý proces s výše uvedenými kroky, a tedy na pojmy dolování dat a získávání znalostí dat je pohlíženo jako na synonyma, k čemuž se také přiklání Han a kol. (2012), ale zároveň souhlasí, že formálně je dolování dat pouze jedním, avšak nejdůležitějším krokem v KDD procesu. S tímto tvrzením se rovněž ztotožnuje autor této práce.

Metody a úlohy dolování dat využívají postupy a techniky z mnoha oborů jako jsou: statistika, databázové technologie, business intelligence, strojové učení, umělá inteligence, neuronové sítě, znalostní systémy a jiné. S využití postupů a technik z těchto oborů pak na dolování dat nahlíží Han a kol. (2012) takto:

Dolování dat je proces získávání znalostí z obrovských množství dat, uložených v databázích, datových skladech, na internetu, nebo jiných informačních uložišťích, či data která proudí systémem dynamicky (Han a kol., 2012, s. 8).

Vztah dolování dat a BI systémů

Novotný, Pour a Slánský (2005, s. 19) definují pojem business intelligence (BI) jako *sadu procesů, aplikací a technologií, jehož cílem je účinně a účelně podporovat rozhodovací procesy ve firmě. Podporují analytické a plánovací činnosti podniků a organizací a jsou postaveny na principech multidimenzionálních pohledů na data.*

Kužela (2015) vidí dolování dat jako jednu samostatnou vrstvu v BI systému (viz Obr. 2.1). Tyto vrstvy znázorňují, v jakém stavu se nacházejí data, resp. informace a jaké techniky jsou v jednotlivých vrstvách použity pro uchovávání a zpracování. Šipka na levé straně obrázku znázorňuje zvyšující se význam pro podporu rozhodování napříč vrstvami. V pravé straně obrázku jsou zmíněny profesní odbornosti lidí, kteří přicházejí s BI systémem do styku v rámci jednotlivých vrstev.

Obrázek 2.1 - Model vrstev BI systému (zdroj: Kužela, 2015)

2.1.2 Případy užití metod vyhledávání asociačních pravidel

Vercellis (2009) mluví o těchto možných případů užití metod vyhledávání asociačních pravidel:

- **Analýza nákupního košíku** – jedná se o analýzu transakcí v maloobchodní síti. Jakmile zákazník zaplatí na pokladním místě za nějaké zboží, je v informačním systému vytvořena jedna transakce. Zaznamenané transakce v databázi jsou pak podrobeny analýze z hlediska přítomnosti opakujících se pravidel vztahujících se k jednotlivým produktům nebo skupinám produktů. Samotné pravidlo může znít obecně jako: „Zákazník kupující nějaký produkt si také koupí jiný produkt s jistou pravděpodobností.“ Výsledky této analýzy pak slouží zejména pro marketingové oddělení společnosti, kterému pomáhají se lépe rozhodnout o umístění sortimentu v prodejně, či strategii ve tvorbě výhodných akcí na produkty.
- **Web mining** – dolování pravidel na webu se týká aktivity uživatelů v prostředí internetových stránek. Podobně jako v předchozím případě zde figuruje člověk vytvářející jistou „transakci“ klikáním na odkazy webových stránek v určitém časovém intervalu. Analýzou těchto transakcí se poté hledá opakující kombinace rozkliknutých odkazů. Obecný příklad pak zní podobně: „pokud člověk (uživatel) navštíví nějakou

webovou stránku, pak do týdné navštíví také jinou stránku s jistou pravděpodobností“. Tento typ asociačních pravidel může být prospěšný pro webové designery, navrhující strukturu webových stránek (optimalizování navigace mezi stránkami, umístění reklamních bannerů).

- **Platby kreditními kartami –** toto je obdobný příklad jako první uvedený. Zde jsou transakcemi nákupy a platby držitelů platebních karet. Nalezená pravidla pak opět slouží pro marketingové účely. Narozdíl od analýzy nákupního košíku zde může být rozmanitost produktů a služeb podléhajících analýze potenciálně neomezená.
- **Detekce podvodů –** s touto kategorií se lze setkat v podstatě jakémkoli průmyslovém odvětví. Vercellis (2009) uvádí příklad pojišťovacích společností, v tomto případě se transakce skládají z ohlášených incidentů a žádostí na kompenzace za způsobenou újmu. Odhalení pravidel v těchto datech může podnítit podezření na podvodnou činnost, které povede k hlubšímu prošetření člověkem.

2.1.3 Způsoby rozdělení asociačních pravidel

Existuje několik druhů asociačních pravidel, které lze rozdělit podle určitých kritérií. U každého specifického druhu se pak používá jiný přístup pro dolování a případně také rozdílné algoritmy. Dle Vercellis (2009) se asociační pravidla dělí dle těchto kritérií (viz Tab. 2.1):

1.	2.	3.	4.
Druh atributů	**Počet dimenzí**	**Druh hierarchie**	**Sekvenčnost transakcí**
pro asymetrické binární atributy	jednodimenzionální	jednoúrovňová	spojitá v čase
pro symetrické binární atributy	vícedimenzionální	víceúrovňová	sekvenční v čase
pro kategoriální atributy			
pro spojité atributy			

Tabulka 2.1 – Způsoby rozdělení asociačních pravidel

Ad 1. – Nejjednodušší podoba asociačních pravidel pochází z transakčních dat s asymetrickými binárními atributy. Binárními proto, že představují buď přítomnost nebo nepřítomnost dané položky v transakci, což lze reprezentovat binární maticí, kde řádky budou jednotlivé transakce a sloupce všechny dostupné položky. V analýze nákupního košíku lze implicitně předpokládat, že přítomnost položky v transakci má daleko větší váhu než její nepřítomnost (asociační pravidla má tedy smysl sestavovat u přítomných položek) – v tomto případě se jedná o data asymetrická. Opakem jsou symetrická data, kde obě binární hodnoty mají stejný význam – např. pohlaví zákazníka. V jiné situaci se mohou vyskytovat atributy kategoriální – např. místo bydliště nebo úroveň vzdělání. Poslední možností jsou spojité atributy (např. věk), kde je třeba diskretizační technikou dospět ke kategorickým atributům.

Pro aplikování algoritmu pro nalezení silných asociačních pravidel, který je popsán dále v práci, je nutno přikročit k transformaci atributů na binární a asymetrické. A to takovým způsobem, že každému kategoriálnímu atributu je přiřazena množina asymetrických binárních proměnných, jejichž hodnota vyjadřuje buď pravdivost nebo nepravdivost atributu.

Ad 2. – Transakční data často bývají uložena v BI systémech, od čehož se odvíjí rozdělení těchto dat podle počtu logických dimenzí. Hledání asociačních pravidel ve vícedimenzionálním prostředí může odhalit mnohem zajímavější a jinak skrytá pravidla – v analýze nákupního košíku tak lze např. zohlednit i konkrétního zákazníka s jeho věkem či příjmy.

Ad 3. – Z důvodu obrovského množství transakcí bývá někdy téměř nemožné sestavit silná asociační pravidla. Produktová data v datových skladech však bývají velmi často uložena v hierarchiích, které je člení do různých druhů a poddruhů. V případě přílišného datového zředění pak stačí využít vyšší úroveň v hierarchii, což povede ke zmenšení počtu objektů podléhající dolování, čímž se také sníží celková výpočetní náročnost. Nebo naopak bývá příznačné najít silná asociační pravidla napříč úrovněmi v hierarchii. Dolování AP (asociačních pravidel) s hierarchickou strukturou je právě věnována celá kapitola 2.2.

Ad 4. – Dle posledního kritéria jsou asociační pravidla rozdělena podle charakteru transakčních dat na spojitá či sekvenční v čase. Lze tedy také získat AP vztahující pouze k určitému časovému intervalu – v tomto případě sekvenční v čase.

2.1.4 Definice asociačních pravidel v analýze nákupního košíku

Nejzákladnější podoba transakčních dat pro stanovení asociačních pravidel má binární asymetrický a jednodimenzionální charakter. Han a kol. (2012) definuje tento druh asociačních pravidel následovně: vycházíme-li z analýzy nákupního košíku, pak univerzem je množina všech artiklů (položek) dostupných v maloobchodě. Každý artikl pak může být reprezentován booleovskou proměnnou, vyjadřující přítomnost či absenci daného artiklu. Nákupní košík je tedy reprezentován vektorem těchto proměnných. Tyto vektory pak mohou být dále analyzovány pro nalezení opakujících se vzorů, jež představují artikly, které jsou opakovaně asociovány neboli společně nakoupeny. Tyto vzory jsou reprezentovány formou asociačních pravidel. Jako hlavní metriky pro určení významnosti asociačních pravidel jsou používány support (překl.: podpora), confidence (překl.: spolehlivost) a také lift (překl.: zvýšení)[1]. Příklad (2.1) takového pravidla s podporou 2% a spolehlivostí 60% je vidět na následujícím vztahu:

$$artikl_1 \Rightarrow artikl_2 \; [podpora = 2\%, spolehlivost = 60\%] \tag{2.1}$$

Nechť $I = \{I_1, I_2, \; .. \; , I_n\}$ je množina literálů, které označují např. všechny dostupné artikly v maloobchodní síti. Nechť T je množina transakcí v databázi, kde každá transakce T_i je množinou položek tak, že platí pravidlo $T_i \subseteq I$. Dále nechť A a B jsou množiny položek tak, že platí $A \subset I, B \subset I, A \neq \emptyset, B \neq \emptyset, A \cap B = \emptyset$. Transakce T_i obsahuje A nebo B tak, že platí $A \subseteq T_i$ nebo $B \subseteq T_i$. Asociační pravidlo je pak implikace $A \Rightarrow B$. U tohoto pravidla pak lze určit podporu *pod* (2.2) jako procento transakcí v množině všech transakcí T obsahujících sjednocení množin A a B ($A \cup B$), což lze označit také jako pravděpodobnost sjednocení těchto množin $P(A \cup B)$. Pravidlo $A \Rightarrow B$ má spolehlivost *sp* (2.3) v množině transakcí T takovou, že *sp* je procento transakcí obsahující množinu B z podmnožiny transakcí obsahujících množinu A. Spolehlivost je tedy podmíněná pravděpodobnost $P(B|A)$. Z výše uvedeného platí vztahy:

$$pod(A \Rightarrow B) \;\; = \;\; P(A \cup B) \tag{2.2}$$

$$sp(A \Rightarrow B) \;\; = \;\; P(B|A) \tag{2.3}$$

[1] V česky psané odborné literatuře se většinou názvy těchto metrik nepřekládají. V této práci však budou dále používány ekvivalentní české výrazy a související názvy proměnných budou odvozeny právě z nich. Aby nedošlo ke zmatení čtenáře vzhledem k podobnosti s šířeji používanými názvy proměnných, je zde použito více než jedno písmeno u názvů proměnných.

Množina *A* je také nazývána jako předchůdce nebo tělo pravidla a množina *B* jako jeho hlava. Pravidla, která splňují (jsou větší nebo rovna) předem stanovený práh pro podporu a spolehlivost, se nazývají silná pravidla. Frekvence výskytů množiny položek je počet transakcí obsahující danou množinu. Podpora pravidla (*pod*) je známa také jako relativní podpora (*rel_pod*) a naproti tomu synonymum pro frekvenci výskytu je absolutní podpora (*abs_pod*). Pokud tedy relativní podpora množiny *I* překračuje práh minimální podpory (*rel_pod* ≥ *min_rel_pod*) nebo absolutní podpora překračuje práh minimální frekvence výskytu, tj. absolutní podpory (*abs_pod* ≥ *min_abs_pod*), pak se jedná o frekventovanou množinu položek.

Z rovnice (2.3) a z výroků výše lze odvodit tento vztah pro výpočet spolehlivosti:

$$sp(A \Rightarrow B) = P(B|A) = \frac{pod(A \cup B)}{pod(A)} = \frac{abs\_pod(A \cup B)}{abs\_pod(A)} \tag{2.4}$$

Z rovnice (2.4) je pak patrno, že spolehlivost pravidla $A \Rightarrow B$ je snadno odvoditelná z frekvence výskytů *A* a $A \cup B$.

Protože ani silná asociační pravidla nemusí být relevantní pro uživatele (což záleží také na subjektivním posouzení každého jednotlivého uživatele), využívají se i podpůrné objektivní metriky, pomocí nichž je posuzována závislost mezi množinami v pravidle. Snahou je zde ještě zúžit množinu pravidel před prezentováním uživateli. Jak znázorňuje rovnice (2.5) jednou z nich je zvýšení (angl.: lift) *z* jako podíl mezi spolehlivostí pravidla a frekvencí výskytu hlavy pravidla. Zvýšením se nazývá proto, že ohodnocuje stupeň, do jakého výskyt jedné množiny „zvyšuje" výskyt druhé množiny.

$$z(A, B) = \frac{P(A \cup B)}{P(A) \cdot P(B)} = \frac{abs\_pod(A \cup B)}{abs\_pod(A) \cdot abs\_pod(B)} \tag{2.5}$$

Pokud $P(A \cup B) = P(A)P(B)$, pak je výskyt množiny *A* v množně transakcí *T* nezávislý na výskytu množiny *B*. V tomto případě je $z = 1$ a tedy neexistuje žádná závislost mezi množinami. Jestliže $z < 1$, pak výskyt množiny *A* je negativně závislý na výskytu množiny *B,* výskyt jedné množiny vede k absenci druhé množiny. Pokud $z > 1$, pak výskyt množiny *A* je pozitivně závislý na výskyt množiny *B,* výskyt jedné množiny vede k výskytu druhé množiny. Tedy výskyt množiny *A* „zvyšuje" pravděpodobnost výskytu množiny *B* koeficientem hodnoty *z*.

Existují ještě další metriky pro určení závislosti mezi množinami. Mezi často používané patří χ^2, metrika absolutní spolehlivosti (sp_{min}), maximální spolehlivosti (sp_{max}), Kulczynskiho, *kosinus* metrika a jiné.

Mezi klasické korelační metriky patří χ^2 (2.6), jejíž hodnota se vypočítá jako suma ze skutečných a očekávaných četností. Pro získání četností je třeba sestavit kontingenční tabulku o rozměrech 2*2, v jejíchž sloupcích a řádcích bude vždy absence a přítomnost dané množiny v transakci. Příklad takové kontingenční tabulky je uveden níže (Tab. 2.2).

$$\chi^2 = \sum \frac{(skutečná\ četnost - očekávaná\ četnost)^2}{očekávaná\ četnost} \tag{2.6}$$

	A	$\bar{A}$	Σ_A
B	*Počet výskytů* AB	*Počet výskytů* $\bar{A}B$	*Počet výskytů* B
$\bar{B}$	*Počet výskytů* $A\bar{B}$	*Počet výskytů* $\overline{AB}$	*Počet výskytů* $\bar{B}$
Σ_B	*Počet výskytů* A	*Počet výskytů* $\bar{A}$	*Počet transakcí*

Tabulka 2.2 - Kontingenční tabulka pro množiny *A* a *B*

Metrika absolutní spolehlivosti sp_{min} (2.7) vyjadřuje nejmenší hodnotu spolehlivosti ze dvou variant podmíněných pravděpodobností a metrika maximální spolehlivosti sp_{max} (2.8) jejich maximum. Výraz $\max\{P(A), P(B)\}$ vyjadřuje maximum ze dvou hodnot podpory množin A a B.

$$sp_{min}(\mathrm{A}, \mathrm{B}) = \frac{P(A \cup B)}{\max\{P(A), P(B)\}} = \min\{\mathrm{P}(\mathrm{A}|\mathrm{B}), \mathrm{P}(\mathrm{B}|\mathrm{A})\} \tag{2.7}$$

$$sp_{max}(\mathrm{A}, \mathrm{B}) = \max\{\mathrm{P}(\mathrm{A}|\mathrm{B}), \mathrm{P}(\mathrm{B}|\mathrm{A})\} \tag{2.8}$$

Kulczynskiho metrika (2.9), která je pojmenována podle polského matematika, je vlastně průměr dvou variant podmíněných pravděpodností vztahujícím se k množinám položek A a B.

$$Kulc(\mathrm{A},\mathrm{B}) = \frac{1}{2}(\mathrm{P}(\mathrm{A}|\mathrm{B}) + \mathrm{P}(\mathrm{B}|\mathrm{A})) \tag{2.9}$$

Na metriku *kosinus* (2.10) je jinak nahlíženo jako na tzv. harmonizované zvýšení. Rozdílem oproti metrice zvýšení je zde odmocnina ve jmenovateli – a tedy její výsledná hodnota je olivněna pouze podporou množin *A*, *B* a $A \cup B$ (resp. podmíněnými pravděpodobnostmi $\mathrm{P}(\mathrm{A}|\mathrm{B})$ a $\mathrm{P}(\mathrm{B}|\mathrm{A})$) a nikoli už celkovým počtem transakcí. Tato metrika může nabývat hodnot pouze z intervalu $\langle 0|1\rangle$. Čím více se číslo blíží jedničce, tím vyšší je závislost mezi množinami *A* a *B*.

$$kosinus(\mathrm{A},\mathrm{B}) = \frac{P(A \cup B)}{\sqrt{P(A) \cdot P(B)}} = \frac{pod(A \cup B)}{\sqrt{pod(A) \cdot pod(B)}} = \sqrt{P(A|B) \cdot P(B|A)} \tag{2.10}$$

Tak jako metrika *kosinus*, i předchozí tři formule výše (2.7, 2.8, 2.9) mají tyto společné vlastnosti:

- jejich hodnoty jsou odvozeny pouze od $pod(A)$, $pod(B)$ a $pod(A \cup B)$,
- mohou nabývat hodnot pouze z intervalu $\langle 0|1\rangle$.
- jejich hodnoty nejsou ovlivněny tzv. nulovými transakcemi.

Co se týče poslední výše zmíněné vlastnosti, počet nulových transakcí představuje počet, kolikrát se v transakcích obě množiny nevyskytují. Jedná se tedy o hodnotu počtu výskytů $\overline{AB}$ v kontingenční tabulce (Tab. 2.2). Metriky zvýšení a χ^2 jsou právě těmito hodnotami silně ovlivněny, protože typicky je počet nulových transakcí mnohem větší než počet transakcí obsahujících alespoň jednu množinu, (Han a kol., 2012).

Pro určení nejlepší volby z předchozích 4 metrik pro posouzení významnosti pravidla představují Han a kol. (2012) metriku poměr nerovnováhy *PN* (2.11), která ohodnocuje nerovnováhu implikace mezi dvěma množinami v pravidle.

$$PN(\mathrm{A},\mathrm{B}) = \frac{|P(A) - P(B)|}{P(A) + P(B) - P(A \cup B)} \tag{2.11}$$

Výsledná hodnota *PN* je definována jako poměr dvou hodnot – v čitateli je to absolutní hodnota rozdílu podpory množiny *A* a podpory množiny *B* a jmenovatel vyjadřuje podporu

transakcí obsahujících množinu A nebo B. Pokud je $pod(A \Rightarrow B)$ stejná jako $pod(B \Rightarrow A)$, pak je $PN = 0$. Čím větší je rozdíl mezi podporami těchto dvou implikací, tím je hodnota PN větší. Han a kol. (2012) doporučují interpretovat tuto metriku spolu s Kulczynskiho metrikou.

2.1.5 Dolování asociačních pravidel

Podle Han a kol. (2012) je na dolování asociačních pravidel nahlíženo jako na proces o dvou krocích:

1) **Generování frekventovaných množin položek –** tyto množiny musí splňovat minimální absolutní podporu (*min_abs_pod*).
2) **Generování silných asociačních pravidel –** tyto pravidla musí splňovat jak minimální podporu, tak minimální spolehlivost (*min_rel_pod, min_rel_sp*).

Důvodem pro zobecnění dolování na tyto dva kroky je fakt, že k nalezení příslušné hodnoty podpory a spolehlivosti pravidla dostačuje nalézt frekventované množiny položek. Po úspěšném identifikování těchto množin a příslušných hodnot *abs_p* pak stačí použít vzorce uvedené výše.

Ad. 1 – V současnosti je používána řada rozdílných algoritmů a jejich modifikací pro dolování frekventovaných množin. Výběr konkrétního algoritmu záleží na podobě souboru dat, který má podlehnout analýze, a také na požadované podobě výsledných asociačních pravidel. V dnešní době patří mezi nejpoužívanější algoritmy tyto následující:

- Apriori algoritmus
- FP-Growth
- ECLAT

První jmenovaný algoritmus je ze všech nejpoužívanější.[2] Jeho smyslem je nejprve nalezení jistých kandidátních množin, na kterých jsou dále ověřovány skutečnosti zda jsou frekventované v daném souboru dat (transakcí) či nikoli. Jeho principem je využití tzv. apriori vlastnosti (předpokladu):

[2] Tento algoritmus je využíván i v praktické části této práce, a proto je detailněji popsán v následující samostatné podkapitole.

Všechny podmnožiny z frekventované množiny musí být rovněž frekventované (Han a kol., 2012, s. 249).

Algoritmus FP-Growth ("Frequent Pattern Growth" – překl.: růst opakujících se vzorů) byl vivinut jako alternativa k Apriori algoritmu pro zefektivnění dolování u případů, kde vyvstanou problémy s výpočetní výkonem při generování obrovského množství kandidátních množin. Výhoda tohoto algoritmu spočívá v rozložení problému na podproblémy a využití rekurzivního postupu. V průběhu algoritmu jsou data ukládána do stromové struktury ze které jsou následně extrahovány veškeré frekventované množiny.

Při užití předchozích dvou algoritmů se předpokládá, že transakce jsou uloženy ve formátu {*TID* : *mn*}, kde *TID* je ID transakce a *mn* je množina obsahující nakoupené položky v transakci *TID,* čemuž se říká horizontální způsob uložení dat. Algoritmus ECLAT ("Equivalence Class Transformation" – překl.: transformace ekvivalentních tříd) pak v prvním kroku tato data transformuje do vertikálního způsobu uložení {*pol* : *TID_mn*}, kde *pol* je název položky a *TID_mn* je množina ID transakcí obsahující danou položku. Dolování pak dále probíhá nad tímto formátem pomocí hledání průniku množin *TID_mn.*

Ad 2. – Po dokončení jednoho z algoritmů jsou nalezeny všechny frekventované množiny v množině transakcí *T* (splňují *min_pod*), a tedy lze přistoupit k dalšímu kroku dle vzorce (2.4) pro určení hodnoty spolehlivosti *sp*. Jak zmiňuje Han a kol. (2012), toto se děje opět v dalších dvou algoritmizovatelných krocích[3]:

- pro každou frekventovanou množinu *l* generuj všechny neprázdné podmnožiny z *l,*
- pro každou neprázdnou podmnožinu *m* z *l* vrať pravidlo $m \Rightarrow (l - m)$, pokud $\frac{min\_abs\_pod(l)}{min\_abs\_pod(m)} \geq min\_sp.$

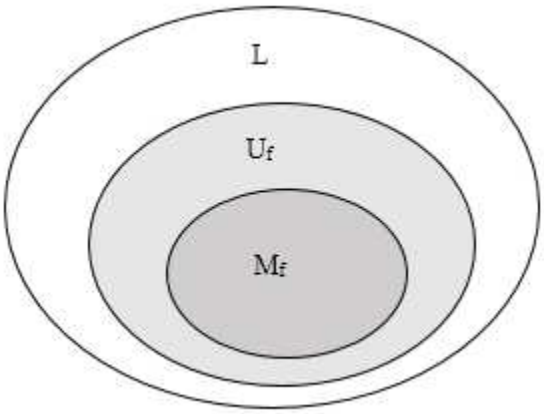
V průběhu celého procesu dolování pak bývá nežádoucím jevem výpočetní náročnost, která u prvního kroku s počtem nalezených transakcí *T* splňující podmínku *abs_pod* ≥ *min_abs_pod* roste. Toto může být problém při obrovském množství dat podlehající analýze a zvláště pak pokud jsou hodnoty *min_abs_pod* nastaveny nízko. Druhý krok se již odvíjí od kroku prvního a dá se tedy říci, že celková výpočetní náročnost je determinovaná výpočetní

[3] Množiny jsou označeny malými písmeny v souvislosti s notací u následující podkapitoly.

náročností prvního kroku. Těmto problémům lze mimo jiné předejít tak, že se dolování omezí na nalezení tzv. uzavřených frekventovaných množin a maximálních frekventovaných množin (Han a kol., 2012):

Uzevřená frekventovaná množina U_f je taková množina v množině transakcí T, že neexistuje žádná nadmnožina X, která by měla stejnou hodnotu absolutní podpory jako U_f v T a zároveň U_f je frekventovanou množinou v množině transakcí T.

Maximální frekventovaná množina M_f je taková množina v množině transakcí T, že neexistuje žádná nadmnožina X, kde platí $M_f \subset X$ a X je frekventovaná množina v T.



Obrázek 2.2 - Vztah frekventovaných množin

Význam množiny U_f spočívá v tom, že každá její podmnožina i prvek má stejnou hodnotu *abs_pod,* z čehož vyplývá že množina U_f nese kompletní informaci o všech jejích podmnožinách. U maximální frekventované množiny M_f se dá pouze říci, že už nelze nalézt další nadmnožinu, která by byla frekventovaná v T. Vztah mezi těmito množinami popisuje také obrázek (Obr. 2.2) výše, kde obsah útvarů znázorňuje počet maximálních, uzavřených a frekventovaných množin v T.

2.1.6 Apriori algoritmus

Apriori vlastnost, jak je popsána výše, je využita dle Han a kol. (2012) v hledání frekventovaných množin ve dvou následovných krocích:

a) Křížové spojování

Principem je nalezení množiny L_k, což je množina všech frekventovaných množin v množině transakcí T, kde k je počet prvků v množině. Pro nalezení této množiny je vygenerovaná kandidátní množina C_k jako kartézský součin dvou množin (2.12).

$$C_k = L_{k-1} \times L_{k-1} \tag{2.12}$$

b) Ořezávání

Množina C_k je nadmnožinou L_k, obsahující jak frekventované tak nefrekventované množiny. Prohledáním všech transakcí lze determinovat *abs_p* každé množiny v C_k. Množiny, které splňují pravidlo *abs_p* ≥ *min_abs_p*, pak tvoří prvky množiny L_k. Apriori vlastnost tedy znamená, že pokud jakákoli podmnožina o *k-1* prvcích z kandidátní množiny C_k není též v L_{k-1}, pak tato množina může být z C_k vyřazena. Toto testování podmnožin se zpravidla děje pomocí zaznamenávání frekventovaných množin do hash stromu, (Han a kol. 2012).

Apriori algoritmus zapsaný pomocí pseudokódu vypadá následovně:

```
(1)   L1 = find_frequent_1-itemsets(T); //prvotní nalezení frekventovaných jednoprvkových množin
(2)   for (k = 2; L_{k-1} ≠ ∅; k++) {
(3)       C_k = apriori_gen(L_{k-1});
(4)       for each transaction T_i ∈ T { // skenování množiny T pro nalezení počtu výskytů
(5)           C_t = subset(C_k, t); // funkce subset vrátí podmožiny t, které jsou kandidátní
množiny
(6)           for each candidate c ∈ C_t
(7)               c.count++;
(8)       }
(9)       L_k = {c ∈ C_k | c.count ≥ min_sup}
(10)  }
(11)  return L = ∪_k L_k;

procedure apriori_gen(L_{k-1}:frequent (k − 1)-itemsets)
(1)  for each itemset l_1 ∈ L_{k-1}
(2)    for each itemset l_2 ∈ L_{k-2}
(3)      if (l_1[1] = l_2[1])∧(l_1[2] = l_2[2])∧ ... ∧(l_1[k − 2] = l_2[k − 2])∧(l_1[k − 1] = l_2[k − 1]) then {
(4)          c = l_1 ⋈ l_2; // krok křížového spojování – generování kandidátů
(5)          if has infrequent subset(c, L_{k-1}) then
(6)              delete c; // krok ořezávání – zbavení se nefrekventovaných kandidátních množin
(7)          else add c to C_k;
(8)      }
(9)      return C_k;

procedure has_infrequent_subset(c: candidate k-itemset;
              L_{k-1}: frequent (k − 1)-itemsets); // použití předchozí znalosti
(1)      for each (k − 1)-subset s of c
(2)          if s ∉ L_{k-1} then
(3)              return TRUE;
(4)      return FALSE;
```

Obrázek 2.3 - Pseudokód pro Apriori algoritmus (zdroj: Han a kol., 2012)

Vstupem pro tento algoritmus je množina transakcí *T* a hodnota minimální absolutní podpory *min_abs_p*. Výstupem je množina frekventovaných množin *L* z *T*. V pseudokódu jak jej definuje Han a kol. (2012) jsou ponechány všechny názvy v anglickém jazyce, tak aby byla zachována nepsaná konvence o psaní programového kódu. V komentářích za dvěmi lomítky jsou pak některé klíčové řádky vysvětleny v češtině.

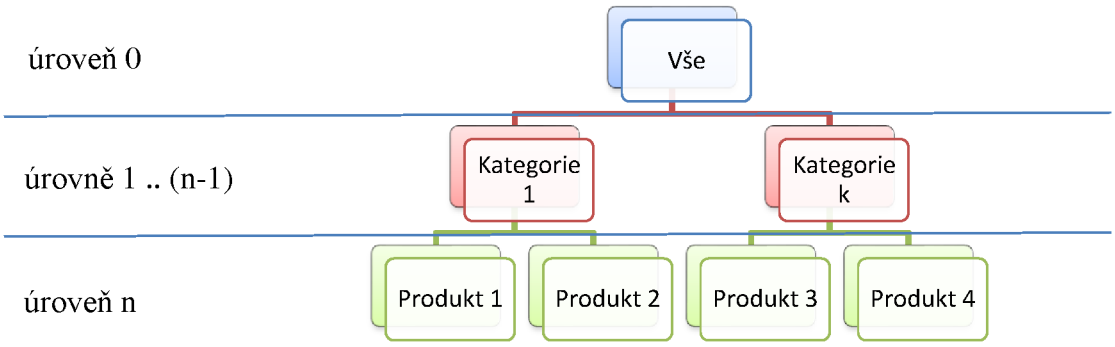
Prvních 11 řádků algoritmu tvoří hlavní kostru algoritmu, na jehož konci jsou nalezeny všechny frekventované množiny v *T*. V průběhu toho je volána procedura *apriori_gen*, která provádí oba základní kroky algoritmu, jak je popsáno výše. Uvnitř této procedury je volána další procedura *has_infrequent_subset*, která vrací booleovskou hodnotu podle toho, jestli aktuální kandidátní množina má nefrekventovanou podmnožinu či nikoli.

2.2 Dolování asociačních pravidel s hierarchickou strukturou

V úvodní podkapitole bude nejprve definován tento druh asociačních pravidel a objasněny přístupy k jejich dolování. V další části této kapitoly bude pojednáno o volbě minimální podpory a spolehlivosti. V poslední části budou stručně uvedeny různé současné algoritmy pro dolování asociačních pravidel s hierarchickou strukturou.

2.2.1 Formulace problému dolování asociačních pravidel s hierarchickou strukturou

Jak již bylo zmíněno v kapitole 2.1.3, asociační pravidla lze rozdělit dle různých kritérií, z nichž jedním je způsob uložení položek (produktů) v databázi – s využitím hierarchické struktury či bez jejího využití. Takto uložená data lze vizuálně zobrazit jako strom hierarchií (Obr. 2.4).



Obrázek 2.4 – Hierarchická struktura produktů

Obecně lze říci, že produktová databáze může nabývat n úrovní, kde každá úroveň obsahuje k kategorií ($n, k \in N$). První úroveň zahrnující všechny podúrovně se nazývá nultá úroveň. Takto bývají uložena v databázi nominální data atributů produktová skupina a produkt. Numerická data lze též rozdělit do hierarchické struktury pomocí diskretizačních technik. Existuje však také možnost, aby si uživatel sám specifikoval podobu hierarchie (Han a kol., 2012).

V rámci řízení kategorií v maloobchodu jsou v praxi zpravidla produkty zařazeny do 2 až 5 úrovní podle stanoveného plánu řízení kategorií (viz kapitola 2.4.2).

2.2.2 Volba minimální podpory a spolehlivosti

Pro volbu minimální podpory a spolehlivosti neexistuje objektivní metoda (nejen u AP s hierarchickou strukturou), podle které se lze za každé situace řídit. Tato volba tedy závisí na subjektivním posouzení potřeb uživatelem. Han a kol. (2012) říká, že se obecně při dolování tohoto typu AP využívá postup „od shora dolů", kde se postupně pro každou úroveň aplikuje jakýkoli algoritmus pro hledání opakujících se vzorů. V tomto postupu existuje několik přístupů jak zvolit hodnotu minimální podpory:

a) **Užití stejné minimální podpory pro všechny úrovně –** výhodou tohoto přístupu je, že stačí, aby uživatelé definovali pouze jednu minimální podporu. V algoritmu v tomto případě může být využito apriori vlastnosti tak, že lze počítat s tím, že každý předek v hierarchii je nadmnožinou jeho potomků. Nevýhodou je však fakt, že na nižší úrovni abstrakce se položky zpravidla budou vyskytovat méně často a může se tak stát, že na výstupu algoritmu budou vynechány některé významné asociace. Naopak při zvolení příliš nízké minimální podpory může zapříčinit vygenerování mnoha nezajímavých asociací.

b) **Užití menší minimální podpory na nižších úrovních –** nevýhody prvního přístupu jsou odstraněny tím, že pro každou úroveň je zvolena specifická minimální podpora tak, že na nižších úrovních je vždy menší než nad jejich nadúrovních.

c) **Užití minimální podpory pro jednotlivé položky nebo skupiny položek –** poslední uvedený přístup je založen na důvěře v uživatele nebo experty, kteří vyberou položky či skupiny položek, pro které bude zvolena minimální podpora individuálně podle kritérií, jimiž může být např. cena položky nebo její subjektivně posouzená významnost. Po identifikování takovýchto položek, resp. skupin položek, a zvolení příslušných hodnot minimální podpory se pro dolování ostatních položek zpravidla využívá nejmenší doposud zvolená minimální podpora.

Je rozdílem pokud dolování pravidel probíhá v rámci každé úrovně zvlášť, anebo také napříč všemi úrovněmi. V prvním případě lze pouze stanovit podporu a spolehlivost pro jednotlivé úrovně zvlášť a pak pouze odstranit redundantní a identifikovat významná pravidla. Ve druhém případě samotné dolování probíhá podle stejného scénáře jako u základních pravidel právě s využitím přístupů zmíněných výše.

2.2.3 Algoritmy pro dolování asociačních pravidel s hierarchickou strukturou

Z praxe je známo, že pro dolování AP s hierarchickou strukturou z transakcí maloobchodích sítí postačuje i základní Apriori algoritmus, který prohledává do šířky zvolenou podmnožinu databáze transakcí (hovoříme-li o řádu desítek a stovek miliónů transakcí z období maximálně několika měsíců). V posledních letech však objemy dat s rozmachem informačních technologií v různých oborech rapidně narůstají – stále více se hovoří o zpracovávání tzv. Big data[4], a hierarchické struktury dat lze také nalézt v jiných oborech než je maloobchod. V současnosti existuje řada studií s novými přístupy k dolování asociačních pravidel s hierarchickou strkturou. V případě transakcí z maloobchodních sítí sice nelze hovořit o Big data, ale přesto tyto studie poukazují na další možná rozšíření, která mohou najít využití i u dolování AP pro analýzu nákupního košíku. A. S. Patel a M. Patel (2015) porovnávají několik aktuálně nejvýznamnějších studií (Tab. 2.3) dolování asociačních pravidel s hierarchickou strukturou takto:

Autoři	Metoda
Qinglan a Longzhen (2013)	Neuronová síť SOFM, Clustering
Vidhate a Kulkarni (2014)	MRA (Bayesovská pravděpodobnost)
Gautam a Pardasani (2010)	Modifikace Apriori
Y. Kim a U. Kim (2009)	HASH Metoda, konceptuální hierarchie
Shrivastava a kol. (2010)	FP-strom
Prakash a Vijayakumar (2011)	Víceúrovňová konceptuální hierarchie
Wan a kol. (2008)	FP-Growth (FP-strom), dynamické hierarchie

Tabulka 2.3 – Studie dolování AP v hierarchických strukturách z posledních let

[4] Tento pojem označující „velká data“ se do češtiny zpravidla nepřekládá. Firma Gartner (2015) definuje pojem Big data jako data, která jsou velkoobměmová, vyžadující okamžité a nákladově efektivní zpracování, vysoce různorodá a jejichž velikost je mimo schopnosti zachycovat, spravovat a zpracovávat data běžně používanými softwarovými prostředky v rozumném čase.

Qinglan a Longzhen (2013) ve své studii představuje využití neuronové sítě pro dolování AP v hierarchicke struktuře. Autoři zde rozšiřují algoritmy „ML_T2L1" a „Cumulate" (Srikant a Agrawal, 1995) a představuji tzv. práh interní podpory, který nahrazuje jinak manuálně nastavovaný práh minimální podpory. Pomocí techniky klastrování dat jsou také data pro dolování rozdělena na několik podmnožin s vlastním prahem interní podpory. Na konkrétní případové studii autoři ukazují, že je jejich metoda využitelná jak v dolování na samostatných úrovních hierarchie, tak napříč různými úrovněmi.

Vidhate a Kulkarni (2014) porovnávájí algoritmus Apriori s jejich vlastním algoritmem MRA pro hledání pravidel v hierarchické struktuře. Z výsledků studie vyplývá, že MRA algoritmus je rychlejší než algoritmus Apriori.

Gautam a Pardasani (2010) navrhují modifikaci Apriori algoritmu pro hledání. Jejich algoritmus se nazývá SC-BF a využívá speciální tabulky minimálních podpor pro generování frekventovaných množin. Algoritmus vyhledává frekventované množiny nejprve na nejvyšší úrovni a poté se dostává k potomkům jednotlivých úrovní. Při prohledávání je využito zmenšených hodnot minimální podpory na nižších úrovních v hierarchii. Mezi výhody tohoto algoritmu podle autorů patří: menší počet vysoce kvalitních pravidel přimo po prvním prohledávání dat a seskupování prvků na každé úrovni, které vede k rychlejší práci s obrovským množstvím dat.

Y. Kim a U. Kim (2009) se zabývají dolováním RFID dat. Pro eliminaci duplikovaných AP napříč úrovněmi využívají generalizaci dat. Jejich metoda se zaměřuje také na redukování velkého počtu AP na některých úrovních v hierarchii – a to takovým způsobem, že ne každá specifická úroveň v jednotlivých hierarchiích má své vlastní prahy minimální podpory. Výsledky jejich studie ukazují, že čas pro generování pravidel tak rapidně poklesl.

Shrivastava a kol. (2010) představují modifikaci FP-Growth algoritmu. Navrhovaná metoda využívá tzv. COFI-strom pro zrychlení dolování ve velkých databázích při využití menší operační paměti.

Prakash a Vijayakumar (2011) představují tzv. frekventovaná a nefrekventovaná asociační pravidla. Uživatelé si podle nového modelu mohou zvolit vícero hodnot minimální podpory. Navrhovaný algoritmus je schopen generovat AP v hierarchické struktuře na základě rozdílné výrokové logiky. Pro dolování je použit tzv. koherentní rámec.

Wan a kol. (2008) ve své studii navrhují využití tzv. dynamických hierarchií, které si uživatel zvolí sám – tato studie se právě zabývá analýzou klasických asociačních pravidel a následným vytvořením hieararchií. Navrhovaná metoda využívá modifikaci algoritmu FP-Growth. Pomocí dynamických hierarchií umožňuje různým uživatelům poskytnout různý pohled na data.

2.3 Implementace dolování asociačních pravidel v prostředí R

2.3.1 Programovací jazyk a prostředí R

Programovací jazyk R byl vyvinut jako open-source software zejména pro statistickou analýzu dat. Je to turingovsky kompletní jazyk, který vzešel z komerčně užívaného jazyka S, umožňující lepší práci s velkým množstvími statistických dat. Jedná se o objektově orientovaný jazyk. Existuje rozsáhlá komunita starající se o vývoj jazyka a jeho knihoven (tzv. balíčků). Každý tento balíček obsahuje funkce pro řešení určité třídy problémů a lze jej nahrát přímo z příkazové řádky do R prostředí tak, že proběhne automatické stáhnutí balíčku z centrálního uložiště CRAN.

Vries a Meys (2012) zmiňují tyto výhody jazyka R:

- dostupnost pro různé operační systémy – Windows, Unix, Mac OS,
- existence rozsáhlé komunity – mailing listy, podporované vývojáři R, mnoho diskuzních fór,
- rozšiřitelnost pomocí balíčků,
- propojitelnost s dalšími programovacími jazyky – Fortran, C, C++, Java, Python.

Vries a Meys (2012) dále uvádějí tyto unikátní prvky R:

- práce s vektory – R je vektorově založený jazyk a lze v něm provádět různé aritmetické operace s vektory,
- zpracovávání i jiných než statistických úloh – ačkoli byl tento jazyk původně vyvinut pro zpracovávání statistických dat, v současnosti pomocí něj lze řešit jakoukoli programovací úlohu a nyní je využíván v oborech jako jsou finance, zpracovávání mluvené řeči, genetika, biologie, marketingový výzkum a další,

- spouštění kódu bez kompilátoru – jedná se o interpretovaný jazyk (což znamená snadnější vývoj při psaní kódu, avšak nevýhodou může být o něco pomalejší běh programu, než tomu bývá u kompilovaných jazyků.

Z výše uvedených důvodu je zřejmé, že je tento jazyk zaměřen primárně na statistické úlohy. Mezi uživatelé tohoto jazyka patří v současnosti stále z velké části odborníci s menšími zkušenostmi s programováním, kteří se zabývají statistikou a analýzou dat. Při využití rozsáhlé škály rozšiřujících balíčku s funkcemi se lze při běžném používání téměř zcela vyhnout jakémukoli použití základních řídících příkazů imperativního programování. Většina těchto R balíčků je napsána v jiných jazycích a to právě z důvodu rychlejšího výpočetního výkonu. Samotný programovací jazyk R se tedy příliš nehodí pro přímou implementaci např. právě algoritmů pro dolování asociačních pravidel, u kterých bývá výpočetní náročnost rozhodující.

2.3.2 Dolování asociačních pravidel v prostředí R

V současné době jsou v repozitáři CRAN k dispozici dva základní[5] balíčky[6], týkající se dolování asociačních pravidel, jsou jimi:

1. **arules**
2. **arulesViz**

Ad 1. – Balíček arules obsahuje implementaci algoritmů Apriori a ECLAT v jazyku C a poskytuje prostředí pro reprezentaci, manipulaci a analýzu transakčních dat (z hlediska dolování asociačních pravidel). Balíček obsahuje třídu vlastní třídu *rules* pro nalezená asociační pravidla (Hahsler, 2005). Mezi nevýhody tohoto balíčku patří nemožnost dolování napříč různými úrovněmi v hierarchii a tvorba AP pouze s jednou položkou na pravé straně pravidla. Výhodou je rychlost implementovaných algoritmů.

Ad 2. – Balíček arulesViz vznikl jako rozšíření balíčku arules. Slouží pro vizualizaci asociačních pravidel přimo v prostředí R s možností vygenerování a množin položek z transakčních dat. Pro rozličné možnosti vizualizace jsou v tomto balíčku napsány tyto funkce, které vykreslují grafy ze třídy *rules*:

[5] Dále jsou k dispozici další dva balíčky arulesNBMiner a arulesSequences, v nichž jsou implementovány algoritmy pro specifické problémy dolování AS.
[6] Balíček v jazyku R je kolekce R funkcí, dat a kompilovaného kódu.

- bodový diagram – generuje dvoudimenzionální bodový diagram se zvolenými metrikami významnosti AP, třetí metrika je reprezentována barevným odstínem bodů,
- matice – uspořádává nalezená pravidla do matice, kde na jedné ose jsou množiny z hlavy pravidla a na druhé množiny z jeho těla, metrika významnosti je reprezentována barevným odstínem nebo velikostí sloupce v případě 3D matice,
- skupinová matice – množiny z těla pravidla jsou pomocí metody klastrování uspořádány do „balónků", na osách matice jsou pak opět těla a hlavy pravidel,
- síťový graf – představuje síťový graf mezi všemi množinami v nalezených AP,
- mozaikový graf – reprezentuje samostatné AP jako kontingenční tabulku pomocí dlaždic v obdélníku, která byla vytvořena rekurzivním horizontálním a vertikálním rozdělením, velikost každé dlaždice je úměrná velikosti hodnoty v kontingenční tabulce,
- dvoupatrový graf – obdoba předchozí vizualizace, s tím rozdílem, že je zde využito pouze jedno horizontální rozdělení,
- graf s paralelními souřadnicemi – používá se pro vizualizaci vícedimenzionálních dat, kde každá dimenze je zobrazena zvlášť na ose *x* a osa *y* je sdílená,
- iGraf – v poslední verzi balíčku (1.0-2) se stále jedná o experimentální vizualizaci, která umožňuje interaktivní funkčnost (selekce, zvýrazňování, změnu barev a další) vygenerovaných grafů (bodové grafy a histogramy)

Výsledné vizualizace je též možno pomocí zabudované funkce exportovat jako GraphML soubor, což je speciální druh souboru na bázi XML pro grafy.

2.4 Základní pojmy v problematice maloobchodních řetězců v souvislosti s analýzou nákupního košíku

2.4.1 Maloobchod a typy organizačních formátů

Cimler a Zdražilová (2007, str. 12) definují maloobchod takto:

Maloobchod je podnik (nebo činnost) zahrnující nákup od velkoobchodu nebo od výrobce a jeho prodej bez dalšího zpracování konečnému spotřebiteli. Maloobchod vytváří vhodné seskupení zboží – prodejní sortiment – co do druhů, množství, kvality, cenových poloh – vytváří pohotovou prodejní zásobu, poskytuje informace o zboží, zajišťuje vhodnou formu prodeje a předává marketingové informace dodavatelům.

Cimler a Zdražilová (2007) uvádějí, že typ organizačního formátu maloobchodu závisí na několika různých faktorech, mezi něž se řadí např. druh obchodních aktivit nebo velikost území, na kterém maloobchod působí. Organizační formáty se tedy liší u těchto typů maloobchodu:

- malý nezávislý maloobchod
- větší maloobchod s několika filiálkami[7]
- maloobchodní řetězec
- silně diverzifikovaný maloobchod

V prvním případě jde o maximálně tři úrovně řízení a zaměstnanci jsou nuceni zvládnout velký rozsah úkolů z hlediska specializace. Využívá se zde pouze funkcionální specializace nebo specializace na produktové bázi.

Větší maloobchod pak vyžaduje hlubší specializaci s existencí většího počtu úrovní řízení. U tohoto druhu se už také může vyskytovat funkce manažera kategorie. Pracovník na této funkci je odpovědný za určitou skupinu zboží, definovanou v rámci category managementu (překl.: řízení kategorií).

Maloobchodního řetězec se vyznačuje větší hloubkou specializace s větším počtem funkcionálních oddělení a také případně vyšší mírou standardizace. Pravomoce a odpovědnosti

[7] Filiálka (z lat. filialis – dceřiný) je pobočka obchodní společnosti nebo jiné instituce, která není samostatnou právnickou osobou, (Wikipedie).

jsou zpravidla centralizovány. Filiálky se mohou adaptovat na místní podmínky týkající se nákupních zvyklostí dané oblasti.

Poslední možností je v podstatě síť několika maloobchodních řetězců, které mají jednoho vlastníka, avšak podnikají v různých oborech. V tomto případě je nutná komplexní organizační struktura.

2.4.2 Řízení kategorií

Řízení kategorií je definováno jako proces mezi maloobchodníkem a dodavatelem, podle něhož je řízena kategorie produktů jako samostatná strategická obchodní jednotka. Ta je řízena tak, aby dosahovala dobrých obchodních výsledků se zaměřením na přidanou hodnotu zákazníkovi, (Anderson Consulting, 2000, str. 4).

V rámci řízení kategorií je pak kategorie definována jako samostatně řiditelná skupina produktů nebo služeb, kterou zákazníci vnímají jako provázanou s naplňování jejich potřeb, (Anderson Consulting, 2000. str. 4).

Studie společnosti Anderson Consulting (2000) zmiňuje tyto 4 fáze každodenního řízení kategorií, které se do jisté míry také navzájem prolínají:

a) Definování strategie řízení kategorií

Strategie řízení kategorií by měla být definována každoročně a měla by zahrnovat kompletní analýzu napříč všemi úrovněmi v hierarchii s ohledem na determinování pozice maloobchodníka na trhu a identifikování příležitostí, které na trhu existují. Analýza zahrnuje přiřazení rolí a alokování zdrojů kategoriím. Vstupem v této fázi je celková korporátní mise a strategie, strategie vlastní značky, strategie řízení dodavatelských řetězců a alokace prodejního prostoru.

b) Vývoj plánů kategorií

Po identifikování dodavatelů v první fázi přichází na řadu sestavení plánu kategorie.

Plán kategorií se vyvíjí v šesti fázích:

1. Definování kategorie
2. Popis role kategorie
3. Ohodnocení kategorie

4. Definování výkonnostních metrik
5. Definování strategie kategorie
6. Definování taktiky kategorie

Ad 1. – Tento první krok dpovídá na otázku jak by měl by měly být produkty segmentovány na základě rozhodovacích stromů spotřebitelů, a tedy jaké produkty by měly být zahrnuty v jednotlivých kategoriích. Informace o potřebách spotřebitelů, jejich nákupního chování a zvyklostí jsou obvykle získány mj. z výzkumů, které provedli dodavatelé, nicméně maloobchodníci mohou definovat kategorie i odlišně od dodavatelů. Kategorie by měly být rozděleny alespoň mezi čtyři úrovně.

Ad 2. – Druhý krok zahrnuje přiřazení cílů jednotlivým rolím definovaných v první fázi. Mezi cíle může patřit např. definování firemní image, poskytnutí rovnováhy mezi hodnotu, růstem a ziskem.

Ad 3. – Cílem třetího kroku je posouzení zda se nakupující v rámci kategorie demograficky shodují s cílovými zákazníky celého maloobchodu a determinování segmentů a značek, které jsou pro cílové zákazníky nejdůležitější. Dále je to odhalení klíčových konkurentů v rámci kategorie, měření výkonu segmentů a odhalení produktů, které pomáhají generovat zisky.

Ad 4. – Ve čtvrté fázi jsou kvantifikovány příležitosti na trhu, jenž byly identifikovány během předchozího kroku. Jsou zde definovány ohodnocující metriky jako je hodnota podílu na trhu, celkové prodeje, procento transakcí, které obsahují kategorii, či provedení analýzy nákupního košíku a další.

Ad 5. – Pátá fáze reprezentuje propojení celkové obchodní strategie a strategie jednotlivé kategorie. Cílem je identifikovat, jak by měla celková strategie určovat strategii kategorie a jak strategie kategorie ovlivňuje taktické příležitosti.

Ad 6. – V poslední fázi už přichází na řadu determinování taktiky, pomocí níž bude dosaženo strategických cílů kategorií. Maloobchodník spolu s dodavateli by se měli zaměřit na

určení co nejlepšího souboru taktik, který v konečném důsledku uspokojí jejich cílové spotřebitele z hlediska rozsahu, cenové politiky, promocí a merchandisingu[8].

c) Implementace plánů kategorií

Implementace plánů kategorií je třetí a v mnoha ohledech nejdůležitější fáze v procesu řízení kategorií – je to fáze, ve které se plán realizuje na prodejně. Podstatné aktivity by měly začít již v průběhu plánování kategorií.

d) Zhodnocení výkonu kategorie

V poslední fázi manažer kategorie sleduje výkon kategorií pomocí stanovených metrik. V této fázi může být využíváno také BI systémů, ze kterých jsou pak data pro reporting zobrazována často pomocí KPI indikátorů v přívětivých vizuálních formách, např. pomocí semaforů, které mohou na první pohled indikovat stav dané kategorie. Takto lze efektivně upozornit manažera na potřebu včasného zhodnocení výkonnosti kategorie.

2.4.3 Využití analýzy nákupního košíku

V kapitole 2.1.2. už byla tato analýza stručně představena, jakožto jeden z případů užití metody hledání asociačních pravidel. Jedná se tedy o hojně využívanou metodu pro odhalení nákupního chování spotřebitelů. Pro tuto analýzu se právě nejčastěji využívají algoritmy pro dolování asociačních pravidel.

Larry Gordon (2008) pod záštitou konzultační společnosti The FactPoint Group ze Silicon Valley vypracoval studii s názvem „Jak nejlepší maloobchodníci používají analýzu nákupního košíku pro zvýšení marží a podílu na trhu“. V ní bylo dotazováno přes 50 různě velkých maloobchodníků. Z odpovědí na otázku k čemu maloobchodníci využívají analýzu nákupního košíku jsou ve studii uvedeny tyto případy užití:

- vyvinutí takové propagace a reklam, které povedou k vyšším ziskům,
- přesnější zacílení nabídky zboží,
- vylepšení propagace věrnostních karet průběžnými analýzami,

[8] Barčík (2013, str. 90) definuje merchandising jako komplexní péči o zboží a POP prostředky na místě prodeje, včetně péče o vizuální stránku prodejny.

- přilákání více zákazníků do prodejen,
- zvýšení obsahu a celkové hodnoty průměrného nákupního košíku,
- testování a učení se při využití trhu jako laboratoře,
- podnícení manažerů a nákupčích k chytřejším rozhodnutím,
- přesné determinování cen pro jednotlivé prodejny,
- přizpůsobení skladových zásob existujícím lokálním potřebám.

Na přilákání zákazníků do prodejen má zcela jistě vliv stav nákupního prostředí a nákupní atmosféry. Autoři Cimler a Zdražilová (2007, s. 228) chápou nákupní atmosféru takto:

Nákupní atmosféra je výsledkem působení prostředí maloobchodní jednotky (nákupního prostředí) a jeho kvantitativních i kvalitativních znaků – vlivů na smysly, kdy tyto vlivy jsou částečně vědomé a zčásti podvědomě vnímány jako individuální prožitek.

Dále Cimler a Zdražilová (2007) uvádějí tyto faktory nákupního prostředí, které dle autora této práce mohou být také zohledněny při aplikaci opatření pro podporu prodeje (alespoň většina z nich):

- design prodejny,
- dispoziční řešení prodejny,
- prezentace zboží,
- personál,
- zákazníci.

Design prodejny lze dále rozdělit na vnější a vnitřní design. Vnější design se týká vnějších stimulů, které působí na zákazníka, aby byl vůbec motivován vstoupit do prodejny. Naproti tomu vhodný vnitřní design by měl zpříjemňovat zákazníkovi nákupní atmosféru a tímto jej také podněcovat k tomu, aby se do prodejny rád vracel.

Cimler a Zdražilová (2007, s. 243) definují dispoziční řešení jako *prostorové uspořádání hmotných prvků obchodního provozu v prodejní místnosti*. Rozdílná řešení jsou použity pro různé druhy maloobchodů.

Cílem prezentace zboží je přilákání zákazníkovy pozornosti a podnícení nákupu tohoto zboží. Opět zde existuje několik prezentačních technik, které bývají využity v závislosti na druhu konkrétního maloobchodu. Cimler a Zdražilová (2007) zmiňují tyto techniky:

- vertikální prezentace – zboží stejného druhu je prezentováno ve výstavním zařízení pod sebou,
- horizontální prezentace – zboží stejného druhu je vystavováno vedle sebe (platí zejména pro menší prodejní jednotky),
- otevřená prezentace – předpokládá aktivní zapojení zákazníka,
- tematická prezentace – zboží je prezentováno společně dle tématu, jakým může být např. sezónnost či sportovní událost a další,
- prezentace životního stylu – zboží je segmentováno z hlediska různých možných životních stylů zákazníka dle sociokulturních trendů,
- prezentace příbuzného zboží – je zde snaha o vytvoření komplementárního efektu, který povede zákazníka ke koupi dalšího zboží,
- prezentace v blocích – je vhodná pro vystavení nového zboží, či zboží za speciální cenu.

Celkovou nákupní atmosféru nakonec dotváří poslední, ti nejdůležitější činitelé, jimiž jsou samotný personál prodejny a zákazníci.

3 Návrh řešení pro efektivní tvorbu a aplikaci asociačních pravidel

3.1 Analýza potřeb implementace a podoby dat

Nejmenovaná softwarová společnost se zabývá již řadu let vývojem informačních systémů v různých sférách podnikání. Mezi hlavní oblasti poskytovaných služeb patří vývoj BI řešení pro maloobchodní sítě. Tato řešení jsou z velké části postavena na technologii Oracle a v současné době umožňují také mimo jiné provádět analýzu nákupního košíku formou hledání asociačních pravidel z transakčních dat. Lze tedy říci, že nejčastějšími zákazníky (uživateli) jsou právě maloobchodní řetězce.

3.1.1 Popis procesu dolování asociačních pravidel a analýza potřeb implementace

Formou rozhovoru bylo zjištěno, že softwarová společnost využívá pro hledání asociačních pravidel z transakčních dat maloobchodních sítí R balíček *arules*, jehož součástí je implementace algoritmu Apriori v jazyku C. Pomocí R skriptu s tímto algoritmem je umožněno hledání asociačních pravidel mezi jednotlivými produkty vždy pouze zvlášť na jedné úrovni v hierarchické struktuře uložených produktů.

Uživatel má přístup k asociačním pravidlům skrze několik prostředí, která jsou součástí širšího BI řešení:

- analytický nástroj QlikView
- prostředí Oracle BI
- nástroj R

Uživatelé pracují především s QlikView a Oracle BI. V obou případech se jedná o práci buď přímo v aplikaci, nebo skrze webové rozhraní, kde se uživateli vygeneruje pouze seznam pravidel pro jednotlivé úrovně v hierarchii spolu s hodnotami metrik podpory, spolehlivosti a zvýšení. Oracle BI je pouze zobrazovací nástroj, který generuje asociační pravidla pouze periodicky za 30 dní. V nástroji QlikView si však uživatel může předem vyfiltrovat AP za

jednotlivé prodejny, úrovně v hierarchické struktuře produktů a různě dlouhá časová období. Seznam AP je v tomto případě jedním z několika analytických reportů.

Uživatel má také možnost zvolit si hodnotu minimální podpory a spolehlivosti – to ale jen v případě práce přímo v nástroji R. V ostatních dvou postředích jsou tyto hodnoty nastaveny podle předem nabytých zkušeností. Pokud má však uživatel explicitní požadavek pro tuto či jakoukoli jinou funkčnost, je možno tomu řešení v QlikView nebo Oracle BI přizpůsobit. Prakticky žádné meze (až na technologické) při implementaci dolování AP nejsou kladeny. Do doby konání rozhovoru však žádný z uživatelů toto nepožadoval.

Z rozhovoru dále vyplynulo, že vědomým nedostatkem v tomto procesu je nejednoznačné determinování významných pravidel na odlišných úrovních v hierarchické struktuře produktů a jejich kategorií a také obtíže v porozumění zákazníka nalezeným pravidlům a celému procesu dolování. S ohledem na tento nedostatek je požadováno vytvoření jakési metodické podpory (popis posloupnosti kroků v průběhu dolování) při stávajícím způsobu dolování (na každé úrovni zvlášť s využitím balíčku arules), která by zahrnovala:

- návrh jak lépe determinovat významná pravidla,
- další návrhy možných variant zlepšení procesu dolování,
- popis procesu s důrazem na praktický (podnikatelský) význam jeho částí.

Je také požadováno, aby využití části této metodické podpory (tj. převážně interpretace pravidel) bylo ilustrováno na vzorku reálných, anonymizovaných dat.

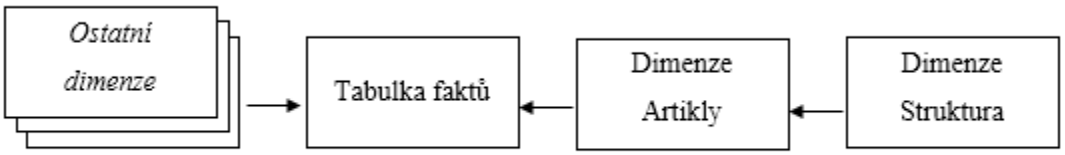
Sekundárním požadavkem je navržení jiného vylepšení procesu bez nároku na jakékoli technologické omezení, který by přinesl přidanou hodnotu pro zákazníka. Jako zřejmý nedostatek pak lze označit nemožnost dolování pravidel napříč (mezi) rozdílnými úrovněmi v hierarchii.

V úvodu proklamovaný cíl práce lze tedy rozdělit na dva vzájemně oddělené podcíle, jimiž jsou:

- zpracování metodické podpory k procesu dolování pomocí R balíčku arules, která bude zahrnovat návrhy jak lépe determinovat významná pravidla a jiné další návrhy na zlepšení procesu a evaluace části tohoto řešení na vzorku anonymizovaných dat,
- návrh jiného zlepšení procesu dolování bez technologických omezení.

3.1.2 Analáza extrahovaných dat

Vyextrahovaná data pocházejí z databáze BI systému, kde jsou rozčleněna podle několika dimenzí. Pro analýzu je podstatná pouze dimenze produktů a struktura produktů (viz Obr. 3.1).



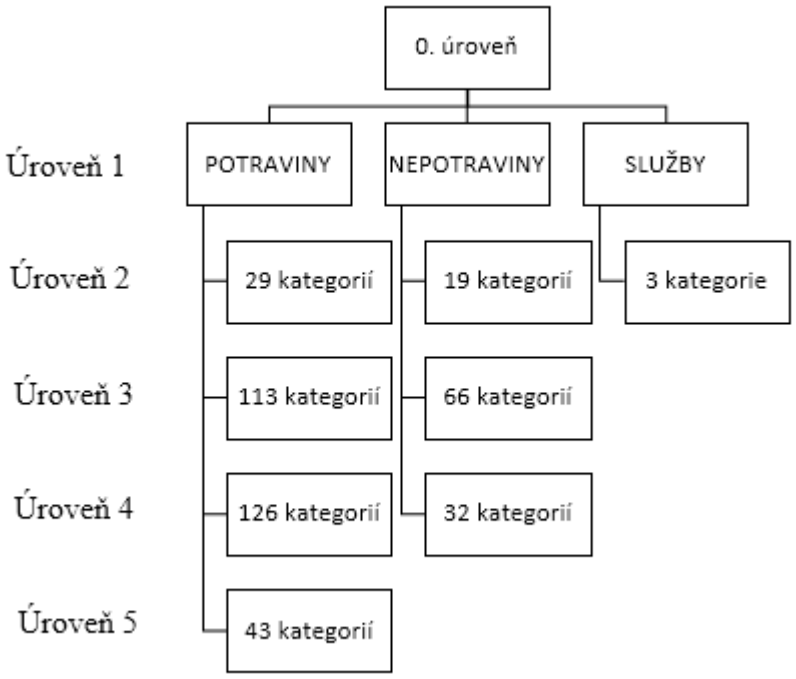
Obrázek 3.1 – Tabulky dat z BI systému

Data pro analýzu jsou uložena v textovém formátu se středníkem jako oddělovačem jednotlivých atributů (CSV formát). Jsou k dispozici celkem tři CSV soubory, z nichž dva soubory jsou dimenze (Artikly, Struktura) a jeden soubor představuje tabulků faktů a obsahuje tedy transakce z maloobchodní sítě. Dimenze Artikly obsahuje veškeré dostupné produkty v prodejnách. V dimenzi Produktová struktura jsou uloženy informace o hierarchické struktuře kategorií pro všechny prodkuty. Tyto dvě dimenze jsou pak propojeny pomocí identifikátoru hierarchické struktury. Tato extrakce byla provedena pouze pro týdenní období. Tabulka (Tab. 3.1) níže obsahuje stručný popis jednotlivých vyextrahovaných tabulek:

	Tabulka faktů	**Dimenze Artikly**	**Dimenze Struktura**
Atributy	Složený PK z CK dimenzí; fakta v podobě prodaného množství a ceny	PK kód artiklu; CK kód struktury; popisné atributy produktů	PK kód struktury; popisné atributy kategorií
1 řádek v tabulce	1 nakoupený produkt	1 produkt	1 kategorie v určité úrovni
Počet řádků	1 290 951	100 518	448

Tabulka 3.1 – Popis vyextrahovaných tabulek

Hierarchická struktura produktů (viz Obr. 3.2) u tohoto vzorku dat obsahuje celkem 6 úrovní, z nichž 5 představují kategorie produktů a v 6. úrovni jsou již samotné produkty. Na obrázku jsou patrny počty kategorií v jednotlivých úrovních pro tři hlavní kategorie. Jak je vidět, ne všechny větve mají stejný počet úrovní, což lze poznat nejen z úplné absence kategorií, ale i z faktu, že některé úrovně mají menší počet kategorií.



Obrázek 3.2 – Hierarchická struktura vzorku dat

3.2 Návrh řešení pro efektivní tvorbu asociačních pravidel při determinaci vztahů mezi produktovými skupinami

Jak vyplynulo z analýzy potřeb implementace, hlavním požadavkem je návrh na zlepšení procesu dolování pomocí R balíčku arules, který zahrnuje navržení vhodné posloupnosti kroků dolování pro mj. snadnější idenifikaci významných pravidel. Tento postup bude spolu s odůvodněními stanoven v první podkapitole. Část tohoto postupu bude dále evaluována na reálných datech v kapitole 4. Druhá podkapitola v této části bude věnována jiným možným zlepšením procesu dolování bez technologického omezení, jež budou posouzena z hlediska přidané hodnoty pro zákazníka a náročnosti případné implementace.

3.2.1 Návrh na zlepšení procesu dolování pomocí R balíčku arules

Předpokladem pro úspěšnost následujícího navrženého postupu je tedy existence funkčního procesu řízení kategorií v maloobchodní síti, dostupnost transakčních dat (z BI systému) a využití statistického nástroje R s balíčkem arules (pro využití Apriori algoritmu).

Po ozřejmení výhod a nevýhod balíčku arules v teoretické části práce a posouzení procesu dolování v analytické části je celý postup při dolování asociačních pravidel v transakcích produktů z maloobchodnách sítí, které jsou uloženy v hierarchické struktuře, navržen v následujících čtyř krocích:

a) Volba rozsahu dolování

Z hlediska rozsahu dolování je příhodné, aby měl uživatel možnost zvolit si tyto parametry přímo v prostředí analytického reportu, které budou použity jako vstup pro algoritmus:

- prodejny,
- období.

Možnost volby prodejen nebo skupin prodejen je vhodná pro rozlišení regionálních nákupních zvyklostí zákazníků.

Pro uživatele může být také přínosná možnost, aby si zvolili specifická data ohraničující délku období pro dolování sami. Uživatelé se tak mohou rozhodnout pro vlastní období na

základě informací o výkonnostech kategorií z reportů BI systému a provést tak zpětně analýzu nad obdobím, které si dle jejich úsudku zasluhuje větší pozornost. Měsíční frekvence dolování zohledňuje sezónnost v chování spotřebitelů a může pomoci nejlépe k určení sezónních marketingových strategií maloobchodního řetězce. Týdenní AP jsou vhodná spíše jen pro zmapování aktuálních prodejních trendů, které následně mohou pomoci zvolit krátkodobou flexibilní marketingovou strategii, tj. na úrovni taktického řízení kategorií. Naproti tomu vydolovaná AP za dlouhá období (> 1 rok) mohou přispět k zkvalitnění rozhodnutí o dlouhodobé strategii maloobchodního řetězce. Je však třeba podotknout, že nalezení kvalitních pravidel může být v tomto velkém počtu transakcí při zvoleném dlouhém období až téměř nemožné. Také nelze opomentout, že více transakcí znamená riziko delší odezvy algoritmu.

Z analytického reportu se tedy odešlou parametry pro selekci příslušných transakcí z databáze, které budou následně vstupem pro R skript. I když jsou navržené parametry voleny z různých dimenzí, stále se pak bude jednat o dolování jednodimenzionálních pravidel – pouze se zúží soubor transakcí určených pro dolování.

b) Provedení analýzy četnosti pravidel pro lepší dolování a stanovení hodnot parametrů pro dolování

Jako podpůrný prostředek k rozhodnutí o zvolení hodnot minimální podpory a spolehlivosti v dalším kroku může posloužit analýza četnosti pravidel. Ta může být řešena algoritmem, který vygeneruje pravidla v intervalech minimální podpory a spolehlivosti zadané uživatelem. Tyto intervaly by mohly být stanoveny pouze jednou před prvním provedením této analýzy. Výsledkem analýzy budou liniové grafy s minimální podporou na ose *x* a počtem pravidel na ose *y*, které lze vykreslit v rámci jednoho reportu v analytickém nástroji. Na jednom grafu lze takto zobrazit sérii křivek při různých hladinách minimální spolehlivosti. Uživatel má možnost z přívětivé vizuální formy posoudit počty pravidel na jednotlivých úrovních při různých minimálních hodnotách a sám se rozhodnout jaké vstupní parametry následně vybrat.

Cílem analýzy je tedy pomoci zvolit takové hodnoty *min_pod* a *min_sp*, aby počet asociačních pravidel byl větší než pro přímou interpretovatelnost všech pravidel, ale ne tak velký, aby znemožňoval jejich další analýzy (viz další krok).

Pro tuto analýzu je potřeba vydolovat všechna pravidla ze zadaných omezení, takže nevýhodou je možné delší trvání algoritmu. V dalším kroku je pak příhodné přistupovat k nalezeným pravidlům přímo do paměti (pokud jsou zvolena kritéria již vydolovaných pravidel).

Jak již bylo uvedeno v kapitole 2.2, existuje několik přístupů pro dolování asociačních pravidel v hierarchické struktuře. Aby však skutečně měla asociační pravidla význam pro uživatele, je velmi důležitá lidská složka, která dokáže nejen nalezená pravidla posoudit, ale také je pomoci sestavit – tedy stanovit hodnotu minimální podpory a spolehlivosti pro jednotlivé úrovně v hieararchické struktuře produktů. Dolování pomocí balíčku arules je omezeno pouze na jednoúrovňová pravidla a tedy je možnost dolovat pouze na každé úrovni zvlášť. S přihlédnutím k analýze četností pravidel tak musí uživatel zvolit minimální hodnoty podpory a spolehlivosti pro jednotlivé úrovně.

Autorem této práce je doporučeno stanovit minimální podporu a spolehlivost tak, aby se počet výsledných pravidel pohyboval v rozmezí asi 300-500 AP na jedné úrovní. Přitom by měla být dodržena strategie volby menší minimální podpory pro menší úrovně. Pro odhalení opravdu zajímavých pravidel je možné využít i jiné metriky pro určení významnosti než je v praxi nejpoužívanější metrika zvýšení. Balíček arules nabízí několik přímo implementovaných metrik, jejich hodnoty pro vydolovaná pravidla lze získat pomocí funkce *interestMeasure()* a propojit je s nalezenými pravidly, které již implicitně obsahují hodnoty podpory, spolehlivosti a zvýšení. Existuje ale také možnost jakoukoli metriku vypočítat přímo z dat. O tom jaké existují přístupy pro zvolení nejvhodnějších metrik pojednává Tan a kol. (2004).

Na konci tohoto kroku je tedy spuštěn dolovací algoritmus a nalezená AP jsou připravena k další analýze a mohou být prezentována v analytickém nástroji, kterým je např. právě QlikView.

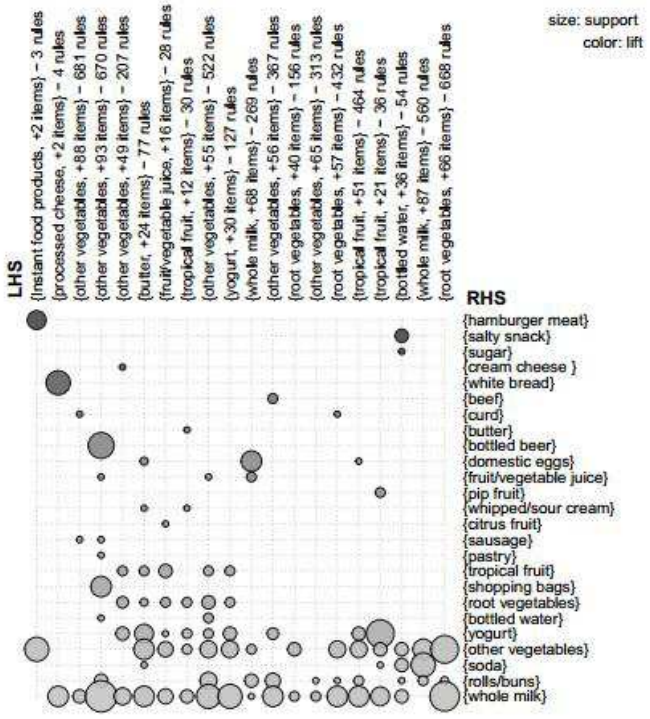
c) Provedení analýzy nalezených asociačních pravidel s použitím metrik významnosti a jejich interpretace

V tomto nejdůležitějším kroku je třeba odhalit zajímavá pravidla pro uživatele. Toto se děje pomocí kategorizace pravidel podle zvolených metrik na:

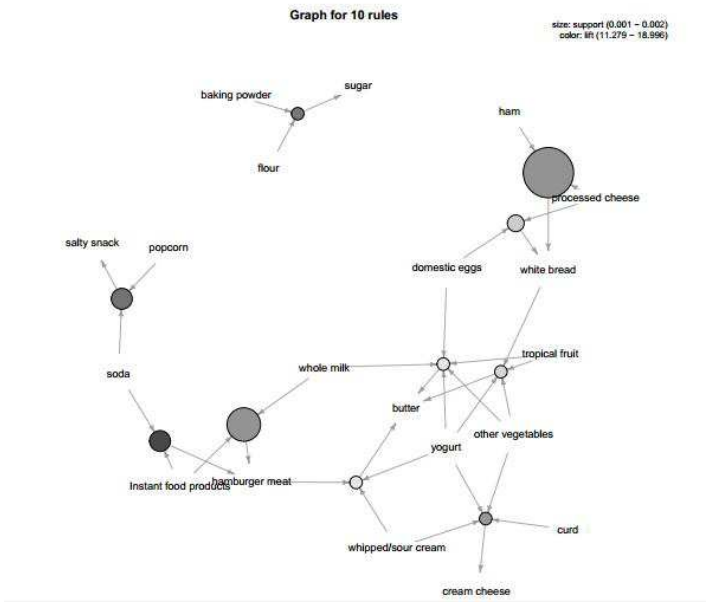
- slabé (neužitečné),
- triviální (uživateli již zřejmé),

- zajímavé pro uživatele.

Klasický tabulkový formát s nalezenými pravidly je vhodný pro zobrazení uživateli, ale pro jednodužší interpretaci pravidel je dobré využít i např. podmíněného formátování při zobrazení tabulkových hodnot jednotlivých metrik. Jako podpůrný prostředek pro interpretaci pravidel je též příhodné využít grafické nástroje nalezených pravidel – např. pomocí příbuzného balíčku arulesViz. Na následujících obrázcích (Obr. 3.3 a Obr. 3.4) je možno vidět příklady dvou vizualizačních technik z balíčku arulesViz – matice AP a síťový graf AP.



Obrázek 3.3 – Matice AP (zdroj: Hahsler, 2005)



Obrázek 3.4 – Síťový graf AP (zdroj: Hahsler, 2005)

Autor této práce navrhuje, aby měl uživatel možnost provádět analýzy sám skrze analytický nástroj, v němž mu bude umožněno stanovovat si omezení pro hodnoty jednotlivých metrik, aby tak zúžil množinu všech zobrazených pravidel. Poté co je uživatel před prvním použitím nástroje důkladně proškolen o významu metrik, může začít samotná analýza. Zajímavou fuknčností by také mohla být možnost vyfiltrování pouze takových existujících pravidel, pro něž zadá uživatel názvy kategorií, resp. produktů. Při takovéto míře samostatnosti dokáže uživatel mnohem snáz odhalit pravidla, která jsou právě pro něj zajímavá.

d) Navržení opatření

Po odhalení zajímavých pravidel, přichází na řadu navržení marketingových opatření na podporu prodeje či přehodnocení strategie kategorie. Způsoby opatření závisí v prvé řadě na rozhodnutí manažera, jehož rozhodování ovlivňuje také obchodní strategie kategorií a jejich aktuální výkonnost. Roli při rozhodování hraje také vynalézavost manažera.

Mezi nejčastěji používaná efektivní opatření z výsledků analýzy nákupního košíku v rámci dolování jednodimenzionálních AP lze označit:

- vytvoření promoční akce na jeden z asociovaných produktů,
- vystavení asociovaných produktů (nebo skupin produktů z dané kategorie) na prodejně ve své blízkosti,

- navržení jiných marketingových opatření s využitím produktových katalogů nebo zákaznických karet,
- a další.

3.2.2 Navržení dalších zlepšení procesu dolování bez technologických omezení

V odborné literatuře je řada příkladů metod, jak naložit s asociačními pravidly v hierarchické struktuře v rámci analýzy nákupního košíku, proto tato kapitola bude věnována rozboru jedné z nich, která se týká proklamovaného nedostatku v kapitole o analýze procesu. Na konci této části bude metoda zhodnocena z hlediska možných přínosů.

Dolování napříč úrovněmi v hierarchii produktů dle Thakur a kol. (2006)

Dle této metody lze řešit nedostatek v procesu dolování, kterým je nemožnost dolování pravidel mezi rozdílnými úrovněmi. Thakur a kol. (2006) ve své práci představuje modifikaci Apriori algoritmu. Metoda autorů je založena na dolování postupně od vyšších úrovní po nižší s využitím redukovaných hodnot minimální podpory a očištěné transakční tabulky na jednotlivých úrovních.

Pro správné použití metody se předpokládá, že v tabulce transakcí je uložena celá informace o větvi hierarchie ve formě posloupnosti číslic. Pořádí těchto číslic znamená konkrétní úroveň a samotná číslice vyjadřuje kategorii na aktuální úrovni. Při provádění algoritmu tento způsob uložení informace znamená také menší nárok na paměť.

Vstupem pro těmito autory navrhovaný algoritmus je tedy tabulka transakcí ve formátu *TID* : *množina*, kde *TID* reprezentuje identifikátor transakce a *množina* obsahuje zakódovanou informaci o hierarchii. Druhým vstupem je hodnota minimální podpory pro každou úroveň v hierarchii. Algoritmus začíná dolování na 1. úrovni, kdy pro každou každou úroveň *l* hledá frekventované množiny z tabulky transakcí *T[l]* o *k* prvcích, jehož výsledkem je tabulka frekventovaných prvků *L[l,k]*.

Thakur a kol. (2006) uvádějí příklad dolování na několia transakcích s položkami z tříúrovňové hierarchie. Na první úrovni jsou tedy nejprve vygenerovány jednoprvkové frekventované množiny do tabulky *L[1,1]* (Obr. 3.6) z tabulky transakcí *T[1]* (Obr. 3.5).

TID	Items
T_1	{111, 121, 211, 221}
T_2	{111, 211, 222, 323}
T_3	{112, 122, 221, 411}
T_4	{111, 121}
T_5	{111,122,211,221, 413}
T_6	{113, 323, 524}
T_7	{131, 231}
T_8	{323, 411, 524, 713}

Obrázek 3.5 – Tabulka transakcí *T[1]* (zdroj: Thakur a kol., 2006)

Itemset	Support
{1**}	7
{2**}	5

Obrázek 3.6 – Tabulka frekventovaných množin *L[1,1]* (zdroj: Thakur a kol., 2006)

Tabulka *L[1,1]* je pak využita pro vygenerování tabulky *T[2]* (Obr. 3.7), očištěné od nefrekventovaných množin. Dále se postupuje vygenerováním tabulky *L[1,2]* (Obr. 3.8) z *T[2]*. Tedy dochází k hledání dvouprvkových množin z transakční tabulky, která byla očištěna o nefrekventované jednoprvkové množiny. Takto se pokračuje dále, dokud nedojde k nalezení všech frekventovaných množin z dané úrovně při zadané hodnotě minimální podpory pro tuto úroveň. V tomto příkladě již byly nalezeny všechny množiny – lze přejít na nižší úroveň.

TID	Items
T_1	{111,121,211,221}
T_2	{111,211,222}
T_3	{112,122,221}
T_4	{111,121}
T_5	{111,122,211,221}
T_6	{113}
T_7	{131,231}

Obrázek 3.7 – Tabulka transakcí *T[2]* (zdroj: Thakur a kol., 2006)

Itemset	Support
{1**,2**}	4

Obrázek 3.8 – Tabulka frekventovaných množin *L[1,2]* (zdroj: Thakur a kol., 2006)

Jak zmiňují Thakur a kol. (2006), důležitým faktem je, že pouze potomci frekventovaných jednoprvkových množin na úrovni 1 mohou být považováni za kandidátní jednoprvkové množiny pro druhou úroveň. Tedy jednoprvkové frekventované množiny pro 2. úroveň *L[2,1]* (Obr. 3.9) jsou nalezeny na základě odvozené transakční tabulky pro jednoprvkové množiny, což je v tomto případě *T[2]*. Následně je *T[2]* očištěna od transakcí, které neobsahují žádnou frekventovanou množinu a tímto je vytvořena *T[3]* (Obr. 3.10). Takto je pokračováno ve všech ostatních úrovních pouze s tím rozdílem, že hledání víceprvkových množin pro další úrovně se děje pomocí párování množin na všech úrovních. Tedy množiny v tabulce *L[2,2]* (Obr. 3.11) jsou nalezeny párováním množin nejen z tabulky *L[2,1]*, ale i z tabulky *L[1,1]*. Přitom alespoň jedna množina pochází z *L[2,1]*.

Itemset	Support
{11*}	6
{12*}	4
{21*}	3
{22*}	4

Obrázek 3.9 – Tabulka frekventovaných množin *L[2,1]* (zdroj: Thakur a kol., 2006)

TID	Items
T_1	{111,121,211,221}
T_2	{111,211,222}
T_3	{112,122,221}
T_4	{111,121}
T_5	{111,122,211,221}
T_6	{113}

Obrázek 3.10 – Tabulka transakcí *T[3]* (zdroj: Thakur a kol., 2006)

Itemset	Support
{11*, 12*}	4
{11*, 21*}	3
{11*, 22*}	4
{12*, 22*}	3
{21*, 22*}	3
{11*, 2**}	4
{12*, 2**}	3
{21*, 1**}	3
{22*, 1**}	4

Obrázek 3.11 – Tabulka frekventovaných množin *L[2,2]* (zdroj: Thakur a kol., 2006)

Obdobně je v navrhovaném algoritmu pokračováno v hledání na všech úrovních, dokud nejsou nalezeny všechny frekventované množiny. V poslední části algoritmu jsou pak derivována asociační pravidla (podobně jako je tomu v klasickém Apriori algortimu) na každé úrovni na základě hodnoty minimální spolehlivosti.

Zhodnocení metody a návrh pro reálné využití
Potřeba pro tento způsob dolování je zřejmá z důvodu možnosti odhalení dalších pravidel, která jinak při dolování na jednotlivých úrovních zůstávají skryta. Při využití dalších vhodných metrik významnosti tak nově nalezená pravidla mohou vést k přesnějšímu rozhodnutí manažerů o aplikaci možných opatření. Pro reálné využití je třeba implementovat vlastní algoritmus. Je možné využít pseudokódu, který je k dispozici v rámci této studie. Volba jazyka pro implementaci by měla být brána s ohledem na dobrý výpočetní výkon. Odhadovaná časová náročnost implementace tohoto algoritmu zkušeným programátorem je v řádu několika týdnů včetně nastudování všech materiálů a testování.

4 Evaluace navrhovaného řešení na anonymizovaných datech

V rámci evaluace řešení bude v první podkapitole popsán postup implementace dolování asociačních pravidel a v následující podkapitole bude provedena evaluace části navrženého řešení na reálných anonymizovaných datech.

4.1 Popis postupu implementace

Pro evaluaci části navrženého postupu na vzorku dat je tedy využit programovací jazyk R, v němž pro potřeby evaluace byly napsány tři skripty.

První skript (viz Příloha č. 1) slouží jako předpříprava dat pro dolování. Probíhá zde transformace vstupních CSV souborů na 6 samostatných CSV souborů ve formátu *TID : název kategorie (název produktu).* Soubory transakce a dimenze jsou načteny do samostatných proměnných se specifickou (pro jazyk R) datovou strukturou – tzv. datových rámců (angl.: data frame), které se používají pro reprezentaci tabulkových dat. Následně je s využitím pomocných proměnných na ně aplikováno několik funkcí tak, že výsledkem je jeden datový rámec, v jehož sloupcích jsou všechny produkty z transakcí spolu s kódem transakce a informacích o všech úrovních v hierarchi, do které daný produkt náleží. Tento datový rámec je pak zdrojem pro rozdělení dat podle úrovní do samostatných datových rámců, které jsou uloženy ve formátu *TID : název kategorie (název produktu)* opět jako CSV soubory.

Druhý skript (viz Příloha č. 2) představuje samotné dolování pomocí balíčku arules a jeho výstup slouží pro analýzu četnosti pravidel navrženou v kapitole 3.2.1. Nejprve jsou načtena data ze souborů (výstup z prvního skriptu) do formátu řídkých matic. Dále je provedeno dolování pro všechny úrovně pomocí funkce *apriori()*. Pro každou úroveň jsou do vektoru uloženy intervaly minimální hodnoty podpory a spolehlivosti a také iterační hodnota pro tyto intervaly. Na základě těchto hodnot pak pomocí cyklů probíhá iterace a jsou vygenerována pravidla pro různé hladiny minimální podpory a spolehlivosti. Pravidla na jednotlivých úrovních jsou uložena do datových rámců a je k nim přidána informace o úrovni. Nakonec jsou všechny rámce spojeny v jeden, který je znovu uložen jako CSV soubor.

Třetí skript (viz Příloha č. 3) je v podstatě jednodušší variantou druhého skriptu s tím rozdílem, že k vydolovaným pravidlům přidává další metriky významnosti pomocí funkce *interestMeasure()*.

Tyto skripty byly napsány pouze pro potřebu zpracování vzorku dat. Ve skutečném procesu dolování by musely být sestaveny s ohledem na co nejpřísnější optimalizaci kódu a načítání dat přímo z databáze BI systému.

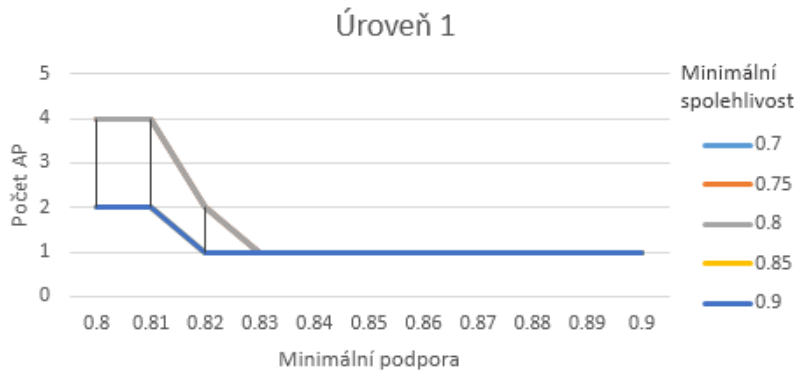
Výstupem z třetího skriptu je CSV soubor, který je pomocí Microsoft Power Query nahrán do Excelu, kde jsou prováděny analýzy v následujících kapitolách.

4.2 Evaluace části navrženého postupu dolování na vzorku dat

Tato kapitola bude rozdělena na dvě podkapitoly, v rámci kterých bude provedena evaluace navrženého řešení na vzorcích dat. V první podkapitole bude nejprve provedena navrhovaná analýza četnosti pravidel a volba parametrů pro dolování a druhá kapitola bude věnována samotné interpretaci pravidel na všech úrovních dolování.

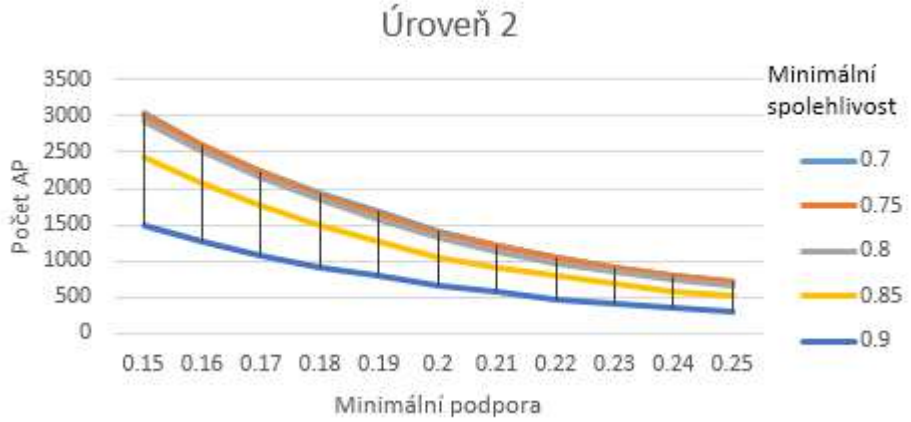
4.2.1 Analýza četnosti pravidel a stanovení hodnot parametrů pro dolování

Pro tuto analýzu byly zvoleny intervaly hodnot minimální podpory a spolehlivosti, které jsou vstupem pro algoritmus, jehož výstupní data slouží pro vykreslení grafů závislosti počtu nalezených pravidel (osa *y*) na minimální podpoře (osa *x*) při různých hladinách minimální spolehlivosti (jednotlivé křivky). Tyto intervaly byly zvoleny s ohledem na čitelnost výsledných grafů. Interval spolehlivosti byl zvolen stejný pro všechny úrovně $min\_sp = \langle 0.7|0.9 \rangle$ a interval podpory je pro každou úroveň rozdílný. Co se týče minimální spolehlivosti, dá se říci, že nemá smysl hledat asociační pravidla s hodnotou $min\_sp < 0.7$, protože při takových hladinách se začínají po dolování vyskytovat pravidla, která lze označit za příliš náhodná.



Graf 4.1 – Četnost pravidel pro 1. úroveň hierarchie

Pro první úroveň (viz Graf 4.1) je zvolen interval minimální podpory velmi vysoko – $min\_pod = \langle 0.8|0.9 \rangle$. Z grafu je patrno, že na dané úrovni se nevyskytují téměř žádná pravidla a ani jich nelze nalézt větší množství, protože se v této úrovni nacházejí pouze 3 kategorie. Dá se tedy říci, že tato úroveň není příliš vhodná pro dolování, anebo je využitelná jen pro dolování triviálních pravidel informativního charakteru. Pro takové dolování tedy ani nemusí být zvolen žádný parametr pro funkci *apriori()* – resp. hodnota parametrů by tak byla nastavena standardně na $min\_pod = 0.1$ a $min\_sp = 0.8$.

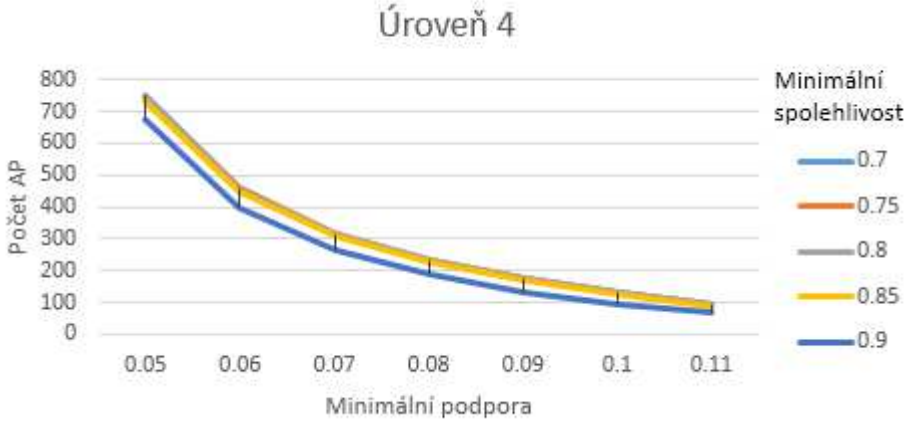


Graf 4.2 – Četnost pravidel pro 2. úroveň hierarchie

Druhá úroveň (viz Graf 4.2) v hierarchii už v daném vzorku transakcí čítá stovky pravidel při nižších hodnotách *min_pod*. V tomto případě byl pro analýzu zvolen interval $min\_pod = \langle 0.15|0.25 \rangle$. K dolování AP pro jejich interpretaci je tedy příznačné zvolit minimální podporu s hodnotou okolo $min\_pod = 0.2$.

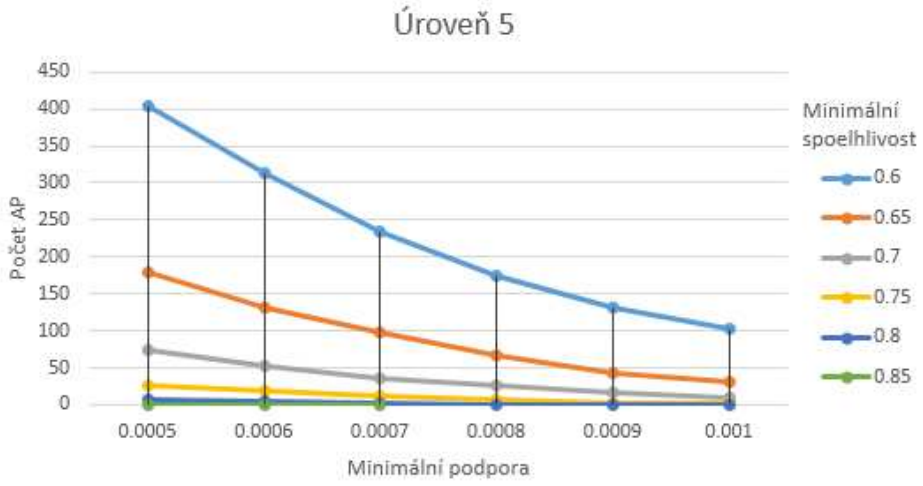
Graf 4.3 – Četnost pravidel pro 3. úroveň hierarchie

Na třetí úrovni (viz Graf 4.3) je vygenerována četnost pravidel v intervalu $min\_pod = \langle 0.1|0.2\rangle$. Z grafu je patrné, že pro dolování je vhodné zvolit $min\_pod = 0.15$. Protože na nižších úrovních minimální podpory se počet AP dostává do řádu tisíců.



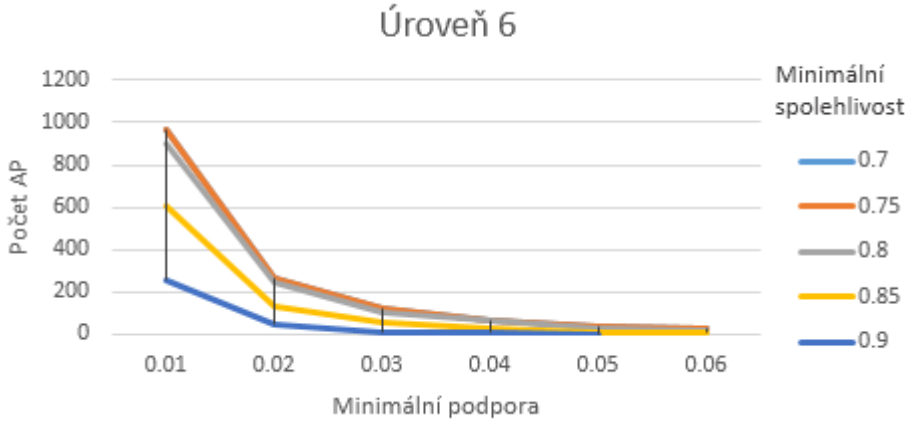
Graf 4.4 – Četnost pravidel pro 4. úroveň hierarchie

Pro úroveň číslo 4 (viz Graf 4.4) je vygenerována četnost pravidel opět pro nižší interval minimální podpory ($min\_pod = \langle 0.05|0.11\rangle$). Z tohoto grafu lze vybrat v podstatě kteroukoli minimální podporu pro dolování, i když nejmenší hodnota min_pod je už poměrně malá a může generovat zbytečně velké množství nezajímavých AP.



Graf 4.5 – Četnost pravidel pro 5. úroveň hierarchie

Pátou úroveň hierarchie (viz Graf 4.5) lze označit za nevhodnou pro dolování, protože se z ní dají vydolovat pravidla pouze při velmi nízkých hodnotách minimální podpory. To je dáno také tím, že úroveň sama o sobě obsahuje velmi málo kategorií.



Graf 4.6 -- Četnost pravidel pro 6. úroveň hierarchie

Poslední úrovní v této analýze je úroveň číslo 6 (viz Graf 4.6), ve které jsou uloženy jednotlivé produkty. Výskyt pravidel je zde opět nevelký i při malých hodnotách minimální podpory. Minimální hodnotu podpory je dobré zvolit okolo $min\_pod = 0.05$.

Po posouzení všech grafů v této analýze je tedy rozhodnuto využít strategii menší minimální podpory na nižších úrovních v hierarchii.

4.2.2 Interpretace pravidel na všech úrovních dolování

Interpretace probíhá podle navrženého postupu v kapitole 3.2.1 – z většího počtu pravidel na každé úrovni jsou hledána významná pravidla s využitím těchto metrik významnosti – *z*, *kosinus* a χ^2 (viz kapitola 2.1.4). V seznamech pravidel jsou na každé úrovni hledána zajímavá pravidla z hodnot dostupných metrik, které jsou podpořeny podmíněným formátováním pro rychlejší orientaci v dlouhých seznamech. V této podkapitole budou postupně představeny všechny úrovně (i ty, které byly označeny v předchozí analýze jako ne příliš vhodné pro dolování) s nejlepšími možnými pravidly a interpretacemi jejich metrik.

a) Úroveň č. 1

Pro dolování na první úrovni jsou použity tyto parametry:

- $min\_pod = 0.1$
- $min\_sp = 0.8$

	Asociační pravidla	*pod*	*sp*	*z*	*kosinus*	χ^2
1	{} => {NEPOTRAVINY}	0.824	0.824	1.000	0.908	NA
2	{} => {POTRAVINY}	0.986	0.986	1.000	0.993	NA
3	{SLUŽBY} => {NEPOTRAVINY}	0.109	0.857	1.040	0.337	45.474
4	{SLUŽBY} => {POTRAVINY}	0.123	0.965	0.979	0.347	183.366
5	{NEPOTRAVINY} => {POTRAVINY}	0.813	0.986	1.001	0.902	7.147
6	{POTRAVINY} => {NEPOTRAVINY}	0.813	0.825	1.001	0.902	7.147
7	{NEPOTRAVINY,SLUŽBY} => {POTRAVINY}	0.108	0.987	1.001	0.328	0.494
8	{POTRAVINY,SLUŽBY} => {NEPOTRAVINY}	0.108	0.876	1.063	0.338	110.041

Tabulka 4.1 – Asociační pravidla na úrovni č. 1

Tabulka výše (Tab. 4.2) představuje všechna nalezená pravidla na první úrovni v hierarchii. Lze je označit za velmi triviální. První dvě pravidla s práznou množinou v těle pravidla znamenají, že bez ohledu na jakékoli jiné produkty v transakcích se množina na pravé straně pravidla objevuje v transakcích s pravděpodobností danou spolehlivostí pravidla (která se rovná podpoře). Tyto první dvě pravidla nemají vypočtenu hodnotu χ^2, protože schází právě jedna množina v pravidle pro vypočtení kontingenční tabulky. Z těchto pravidel se lze pouze

pro informaci dozvědět za jakých pravděpodobností se tyto kategorie vyskytují v transakcích. Zvýšení se u těchto pravidel ve všech případech téměř rovná jedné (kromě prvních dvou kde je hodnota přesně $z = 1$), což znamená, že množiny jsou na sobě nezávislé.

Využitelnost pro manažery je v tomto případě téměř žádná. Dá se předpokládat, že takové informace již mohou získat z jiných částí BI systému.

b) Úroveň č. 2

Pro dolování na druhé úrovni jsou použity tyto parametry:

- $min\_pod = 0.2$
- $min\_sp = 0.8$

Na této úrovni bylo vydolováno celkem 653 pravidel a je velmi těžké odhalit zajímavá pravidla. Většinu z nich lze označit za nezajímavá nebo triviální. Většina pravidel obsahuje v pravé straně pravidla pekárenské a mlékárenské výrobky, které jsou samy o sobě nakupovány ve velké míře, proto se dá předpokládat že se budou v pravidlech nejvíce objevovat. Ty nejzajímavější z nich jsou uvedeny v tabulce (Tab. 4.2) níže.

Číslo AP	Asociační pravidla
1	{MASOVÉ VÝR. ČER.,PEKÁRENSKÉ VÝROBKY} => {MLÉKÁRENSKÉ VÝR.}
2	{NÁPOJE,OVOCE A ZELENINA} => {MLÉKÁRENSKÉ VÝR.}
3	{CUKROVINKY,KUŘÁCKÉ POTREBY} => {PEKÁRENSKÉ VÝROBKY}
4	{KUŘÁCKÉ POTŘEBY,NÁPOJE} => {MLÉKÁRENSKÉ VÝR.}
5	{CUKROVINKY,KUŘÁCKÉ POTŘEBY} => {NÁPOJE}
6	{MLÉKÁRENSKÉ VÝR.} => {PEKÁRENSKÉ VÝR.}
7	{CUKROVINKY} => {PEKÁRENSKÉ VÝR.}

Tabulka 4.2 – Nejzajímavější AP na úrovni č. 2

Číslo AP	*pod*	*sp*	*z*	*kosinus*	χ^2
1	0.608	0.947	1.136	0.831	7047.015
2	0.533	0.946	1.135	0.778	5014.784
3	0.343	0.946	1.116	0.619	1815.458
4	0.355	0.909	1.090	0.622	1110.287
5	0.327	0.902	1.122	0.605	1467.733
6	0.767	0.921	1.085	0.913	8689.855
7	0.689	0.909	1.072	0.860	3840.860

Tabulka 4.3 – Metriky nejzajímavějších AP na úrovni č. 2

Za zmínku stojí některá pravidla s vyšší hodnotou podpory – i když se mohou zdát triviální, některé z nich mohou přinést nový poznatek. Pravidla s menší úrovní podpory už obsahují velké množsví kategorií v těle pravidla a nejsou tak již informačně přínosné. Z tabulky metrik (Tab. 4.3) lze vyčíst, že všechna tyto pravidla mají koeficient zvýšení nepatrně větší než 1. To znamená, že výskyt množin na levé straně pravidla lehce zvyšuje také výskyt množiny na pravé straně pravidla. Vysoké hodnoty metriky χ^2 také indikují závislost mezi množinami na obou stranách pravidel. Hodnota metriky *kosinus* u pravidla č. 6 znamená také velmi vysokou závislost mezi množinami (což lze zpozorovat také z velmi vysoké hodnoty podpory a spolehlivosti).

Trochu zajímavý poznatek mohou přinést pravidla č. 3 a 4, které zahrnují kuřácké potřeby. Ostatní pravidla nejsou příliš zajímavá.

c) Úroveň č. 3

Pro dolování na třetí úrovni jsou použity tyto parametry:

- $min\_pod = 0.15$
- $min\_sp = 0.8$

Na třetí úrovni v bylo vydolováno celkem 584 pravidel. Po posouzení metrik byly vybrány tyto pravidla jako nejzajímavější (Tab. 4.4).

Číslo AP	Asociační pravidla
1	{JOGURTY A JOGURTOVÉ DEZERTY} => {MASOVÉ VÝR. VEPŘ. A HOV.}
2	{PIVO} => {PEČIVO}
3	{MASOVÉ VÝR. VEPŘ. A HOV.} => {PEČIVO}
4	{PEČIVO,SÝRY} => {MASOVÉ VÝR. VEPŘ. A HOV.}

Tabulka 4.4 – Nejzajímavější AP na úrovni č. 3

Číslo AP	*pod*	*sp*	*z*	*kosinus*	χ^2
1	0.455	0.800	1.175	0.731	3644.312
2	0.345	0.858	1.078	0.610	683.156
3	0.613	0.900	1.131	0.833	6061.564
4	0.433	0.849	1.246	0.734	5714.031

Tabulka 4.5 – Metriky nejzajímavějších AP na úrovni č. 3

V této úrovni je obdobný problém jako na úrovni předchozí. Nalezená pravidla jsou spíše nezajímavá a triviální. Z výše uvedených pravidel (Tab. 4.5) lze označit pravidlo č. 2 jako zajímavé, kdy ve více než 30% transakcí se oběvily výrobky z kategorie *pivo* a *pečivo*. Ve více než 80% transakcí s produkty z kategorie pivo se vyskytovalo také pečivo. Hodnoty metrik *kosinus* a χ^2 indikují závislost množin v pravidlech.

d) Úroveň č. 4

Pro dolování na čtvrté úrovni jsou použity tyto parametry:

- $min\_pod = 0.1$
- $min\_sp = 0.8$

Na této úrovni bylo vydolováno celkem 746 pravidel. Následující pravidla byla vybrána jako nejzajímavější (Tab. 4.6).

Číslo AP	Asociační pravidla
1	{CHLÉB NEBALENÝ} => {PEČIVO SLANÉ}
2	{JOGURTY OCHUCENÉ} => {PEČIVO SLANÉ}
3	{OPLATKY} => {PEČIVO SLANÉ}

Tabulka 4.6 – Nejzajímavější AP na úrovni č. 4

Číslo AP	*pod*	*sp*	*z*	*kosinus*	χ^2
1	0.388	0.903	1.171	0.674	3084.140
2	0.385	0.884	1.146	0.664	2309.777
3	0.316	0.870	1.127	0.597	1305.904

Tabulka 4.7 – Metriky nejzajímavějších AP na úrovni č. 4

V této úrovni již všechna pravidla obsahovala v pravé straně pravidla kategorii slané pečivo, které je velmi hojně nakupováno. Pravidla výše byla vybrána na základě největších hodnot podpory a spolehlivosti. Ostatní metriky významnosti (Tab. 4.7) indikují opět podobné informace jako na předchozích úrovních. Je na manažerech maloobchodu, aby posoudili, zda některá z nich jsou zajímavá.

e) Úroveň č. 5

Pro dolování na čtvrté úrovni jsou použity tyto parametry:

- $min\_pod = 0.001$
- $min\_sp = 0.7$

Pátá úroveň obsahuje pouze malý počet kategorií oproti předchozím úrovním. Při zadaných parametrach v ní bylo vydolováno celkem 10 pravidel. Protože však hodnota podpora byla nastavena velmi nízko, nelze očekávat nalezení zajímavých pravidel (v analýze četnosti AP bylo doporučeno na této úrovni ani nedolovat). V tabulce níže (Tab. 4.8) jsou přesto vybrány dvě z nich pro interpretaci metrik.

Číslo AP	Asociační pravidla
1	{PIVO LAHVOVÉ OSTATNÍ,ZNAČKOVÉ LÍHOVINY - TYP A} => {PIVO LAHVOVÉ 10 %}
2	{JOGURTY NEOCHUCENÉ SMETANOVÉ,JOGURTY OCHUCENÉ STŘEDNĚTUČNÉ A PROBIO, MASOVÉ VÝR. DRŮBEŽÍ MĚKKÉ - PÁRKY} => {JOGURTY OCHUTENÉ SMETANOVÉ}

Tabulka 4.8 – Nejzajímavější AP na úrovni č. 5

Číslo AP	*pod*	*sp*	*z*	*kosinus*	χ^2
1	0.001	0.724	2.541	0.054	54.974
2	0.001	0.800	2.166	0.049	39.852

Tabulka 4.9 – Metriky nejzajímavějších AP na úrovni č. 5

Podle metriky *kosinus*, která není závislá na počtu transakcí lze uvést, že množiny v těchto pravidlech na sobě nejsou závislé, přestože metrika zvýšení indikuje pozitivní závislost množin z levé strany pravidla na množinách z pravé strany pravidla.

f) Úroveň č. 6

Pro dolování na šesté úrovni jsou použity tyto parametry:

- $min\_pod = 0.05$
- $min\_sp = 0.7$

Na nejnižší úrovni, která obsahuje samotné produkty bylo vydolováno při zadaných parametrech celkem 35 pravidel. Všechna pravidla obsahují jeden konkrétní produkt v pravé straně pravidla – tento problém už byl částečně odhalen při dolování na vyšších kategoriích. Všechna ostatní pravidla jsou podle metrik obdobně nezajímavá jako následující tři vybraná pravidla (Tab. 4.10).

Číslo AP	Asociační pravidla
1	{BANÁNY} => {SLANÉ PEČIVO 1}
2	{MLÉKO 1} => {SLANÉ PEČIVO 1}
3	{POMARANČ 1} => {SLANÉ PEČIVO 1}

Tabulka 4.10 – Nejzajímavější AP na úrovni č. 6

pod	*sp*	*z*	*kosinus*	χ^2
0.163	0.804	1.181	0.439	760.538
0.169	0.846	1.243	0.458	1335.267
0.117	0.816	1.199	0.374	599.460

Tabulka 4.11 – Metriky nejzajímavějších AP na úrovni č. 6

Podle metrik významnosti (Tab. 4.11) jsou množiny z obou stran pravidel na sobě mírně závislé, ačkoli pro vysoký počet obdobných pravidel lze usuzovat, že jejich informační hodnota je zcela nepřínosná.

Zhodnocení interpretace

Po provedené analýze a interpretaci nalezených pravidel na všech úrovních lze říci, že dolování z poskytnutého vzorku dat bylo silně zkresleno prodejem velkého množství jednoho konkrétního produktu, který byl zřejmě v promoční akci v daném období. Toto je ale pouze domněnka autora této práce, kterou nelze prokázat. Při analýze proběhl také pokus o snížení hodnot parametrů pro dolování, který však vedl ke stejným výsledkům. Během této interpretace tak byla alespoň rozebrána některá pravidla z hlediska možných interpretací zvolených metrik. Na žádné úrovni tedy nebylo nalezeno opravdu zajímavé pravidlo, na základě něhož by mohla být navržena opatření k podpoře prodeje. Je pouze na posouzení manažerů, zda-li je některé z uvedených pravidel alespoň částečně informačně přínosné. K důkladnějšímu posouzení nalezených pravidel je tedy také potřeba znát strategii a taktiku v řízení kategorií daného maloobchodu.

5 Závěr

Tato diplomová práce byla zaměřena na dolování asociačních pravidel s hierarchickou strukturou v maloobchodních sítích. Hlavním cílem práce byl návrh pro zlepšení procesu dolování jednodimenzionálních asociačních pravidel na transakčních datech s hierarchickými strukturami produktů z maloobchodních sítí, které povede k mj. uživatelsky snadnější identifikaci významných pravidel. Pro dosažení tohoto cíle byl využit statistický nástroj R, který byl zároveň podmínkou pro splnění jednoho podcíle práce, v kombinaci s Microsoft Excelem s analytickými doplňky pro evaluaci části navrhovaného řešení.

První část diplomové práce tvoří souhrn teoretických poznatků potřebných pro dosažení cíle práce. Nejprve byly popsány nezbytné základní pojmy týkající techniky dolování asociačních pravidel v rámci metody dolování dat. Dále byla tato technika podrobněji rozebrána s podporou matematického aparátu. Bylo pojednáno také o možných aplikacích dolování asociačních pravidel a různých algoritmech s důrazem na nejpoužívanější algoritmus Apriori. Následně byla ozřejmena problematika asociačních pravidel s hierarchickou strukturou a byla popsána její specifika. Dále nebyl opomenut ani statistický nástroj R s popisem jeho využití k dolování asociačních pravidel. Na konci teoretické části byly také definovány některé pojmy v problematice maloobchodu a bylo pojednáno o využití analýzy nákupního košíku.

Další částí je třetí kapitola, která je věnována návrhu řešení s ohledem na splnění cíle práce. V této části byla provedena analýza stávajícího procesu dolování. Na základě odhalených nedostatků v procesu a přímých požadavků byly definovány dva podcíle práce. Pro splnění prvního podcíle byla dále navržena metodická podpora (popis postupných kroků při dolování), která zahrnuje návrhy na možná vylepšení procesu s důrazem na snadnější odhalení významných pravidel. Část navrhované metodické podpory byla také ověřena na reálných prodejních datech v další samostatné kapitole. Pro splnění druhého podcíle byla také v kapitole návrhu řešení rozebrána studie o jiném algoritmu, pomocí kterého lze řešit uvedený nedostatek v procesu dolování. Tento algoritmus byl také stručně zhodnocen z hlediska nových přínosů pro proces.

Z výsledků evaluace části navrženého řešení bylo zjištěno, že testovaný vzorek dat obsahoval zkreslená data, jejichž příčinu existence nelze přesně prokázat. Přesto se dá říci, že došlo k jistým zlepšením procesu – hlavně z hlediska přehlednosti interpretovaných pravidel po aplikaci podmíněných formátování buněk v tabulkách reportů s nalezenými pravidly a také

bylo poukázáno na možnost využití dalších metrik významnosti. K přesnějšímu rozhodnutí o volbě parametrů pro dolování může také dopomoci analýza četnosti pravidel.

Jako námět pro další práci na tomto problému se tedy nabízí navrhnout další evaluace řešení na jiných vzorcích dat a využití také i dalších metrik významnosti a do této evaluace zapojit také zákazníky, kteří budou asociační pravidla využívat k podpoře rozhodování. Po nalezení významných pravidel a aplikaci příslušných opatření by bylo dobré také vypracovat případovou studii, podle které by byla posouzena účinnost těchto opatření.

Seznam použité literatury

Knižní publikace

BARČÍK, Tomáš. *Strategický marketing*. Praha: Ústav práva a právní vědy, 2013. 110 s. Praha: Ústav práva a právní vědy. ISBN 978-80-905247-7-4.

CIMLER, Petr a Dana ZADRAŽILOVÁ. *Retail management*. Praha: Management Press, 2007. 307 s. ISBN 978-80-7261-167-6.

DE VRIES, Andrie and Joris MEYS. *R for dummies*. Chichester: John Wiley and Sons, c2012, xviii, 387 s. For dummies. ISBN 978-1-119-96284-7.

HAMMOND, Richard. *Chytře vedená prodejna: jak mít více zákazníků a větší tržby*. 2. české vyd. Praha: Grada, 2012, 196 s. ISBN 978-80-247-4162-8.

HAN, J., M. KAMBER and J. PEI. *Data Mining: Concepts and Techniques.* Boston: Elsevier, 2012. 703 s. ISBN 978-0-12-381479-1.

HORÁKOVÁ, Helena. *Marketingové strategie*. 1. vyd. Praha: Idea servis, 2014. 103 s. ISBN 978-80-85970-81-4.

JAKUBÍKOVÁ, Dagmar. *Strategický marketing: strategie a trendy*. 2., rozš. vyd. Praha: Grada, 2013. 362 s. ISBN 978-80-247-4670-8.

LACKO, Luboslav. *Databáze: datové sklady, OLAP a dolování dat s příklady v Microsoft SQL Serveru a Oracle.* 1. vyd. Brno: Computer Press, 2003. 486 s. ISBN 80-722-6969-0.

NOVOTNÝ, O., J. POUR a D. SLÁNSKÝ. *Business Intelligence. Jak využít bohatství ve vašich datech.* 1. vyd. Praha: Grada Publishing, 2005. 254 s. ISBN 80-247-1094-3.

STOWELL, Sarah. *Using R for Statistics.* New York: SPRINGLER, 2014. 244 s. ISBN 978-1-4842-0140-4.

TAN, P., M. STEINBACH and V. KUMAR. *Introduction to data mining*. Vyd. 1. Boston: Pearson Addison Wesley, 2006. 769 s. ISBN 978-0321321367.

VERCELLIS, Carlo. *Business Intelligence – Data Mining and Optimization for Decision Making.* Indianapolis: John Wiley & Sons, 2009. 417 s. ISBN 978-0-470-51138-1.

WITTEN, I., E. FRANK and M. A. HALL. *Data mining: practical machine learning tools and techniques*. 3. vyd. Amsterdam: Morgan Kaufmann, 2011. 629 s. ISBN 978-0-12-374856-0.

ZAKI, Mohammed J. and Wagner MEIRA. *Data mining and analysis: fundamental concepts and algorithms*. New York: Cambridge University Press, 2014. 593 s. ISBN 978-0-521-76633-3.

Články v odborných časopisech a sbornících z konferencí

AGRAWAL, R., T. IMIELIŃSKI and A. SWAMI. Mining association rules between sets of items in large databases. In: *Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93*. New York, New York, USA: ACM Press, 1993, s. 207-216. ISBN 0-89791-592-5.

GAUTAM, Pratima and K. R. PARDASANI. Algorithm for Efficient Multilevel Association Rule Mining. *International Journal on Computer Science and Engineering*. 2010, č. 2, s. 1700-1704. ISSN 0975-3397.

HAHSLER, M., B. GRÜN and K. HORNIK. arules – A Computational Environment for Mining Association Rules and Frequent Item Sets. *Journal of Statistical Software* [online]. 2005, roč. 14, č. 15. [cit. 2015-08-02]. ISSN 1548-7660. Dostupné z: http://www.jstatsoft.org/v14/i15/paper

JIAWEI, Han and Yongjian FU. Discovery of Multiple-Level Association Rules from Large Databases. In: DAYAL U., P. M. D. GRAY, AND S. NISHIO, eds. *Proceedings of the 21th International Conference on Very Large Data Bases - VLDB '95*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, s. 420-431. ISBN 1-55860-379-4.

KIM, Younghee and Ungmo KIM. Mining Multilevel Association Rules on RFID data. In: *First Asian Conference on Intelligent Information and Database Systems.* Dong Hoi, Vietnam: IEEE Computer Society, 2009, s. 46-50. ISBN 978-0-7695-3580-7.

PATEL, Alisha S. and Mohit PATEL. Effective and Innovative Approaches for Comparing Different Multilevel Association Rule Mining for Feature Extraction: A Review. *International Journal of Computer Applications.* 2015, č. 109, s. 46-49. ISSN 0975-8887.

PRAKASH, S., M.VIJAYAKUMAR and R.M.S.PARVATHI. A Novel Method of Mining Association Rule with Multilevel Concept Hierarchy. *International Journal of Computer Applications*. 2011, č. 2, s. 26-29. ISSN 0975-8887.

QINGLAN, Huang and Duan LONGZHEN, Multi-level Association Rule Mining Based on Clustering Partition. In: *Proceedings of the 2013 Third International Conference on Intelligent System Design and Engineering Applications - ISDEA '13.* Washington, DC, USA: IEEE Computer Society, 2013, s. 982-985. ISBN 978-0-7695-4923-1.

SHRIVASTAVA, A., R. C. JAIN and A. K. SHRIVASTAVA. Comparison of New Multilevel Association Rule Algorithm with MAFIA. *International Journal of Intelligent Systems and Applications.* 2014, č. 11, s. 75-81. ISSN 2074-9058.

SHRIVASTAVA, V.K., P. KUMAR and K. R. PARDASANI. Discovery of Multi-level Association Rules from Primitive Level Frequent Patterns Tree. *International Journal of Computing Science and Communication Technologies.* 2010, roč. 3, č. 1, s. 506-510. ISSN 0974-3375.

SRIKANT, Ramakrishnan and Rakesh AGRAWAL. Mining Generalized Association Rules. In: DAYAL U., P. M. D. GRAY a S. NISHIO, eds. *Proceedings of the 21th International Conference on Very Large Data Bases - VLDB '95*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, s. 407-419. ISBN 1-55860-379-4.

SRIVANSTAVA, Arpna and R.C. Jain. Performance Analysis of Modified Algorithm for Finding Multilevel Association Rules. *Computer Science & Engineering: An International Journal (CSEIJ)* [online]. 2013, roč. 3, č. 4 [cit. 20-3-2015]. ISSN 2231 - 329X. Dostupné z: http://arxiv.org/ftp/arxiv/papers/1309/1309.2371.pdf

THAKUR. R.S., R.C. JAIN and K.R. PARDASANI. Mining Level-Crossing Association Rules from Large Databases. *Journal of Computer Science.* 2006, č. 2, s. 76-81. ISSN 1549-3636.

VIDHATE, Deepak and Parag KULKARNI. To improve Association Rule Mining using New Technique: Multilevel Relationship Algorithm towards Cooperative Learning. In: *Circuits, Systems, Communication and Information Technology Applications.* Bombaj, Indie: IEEE Computer Society, 2014, s. 241-246. ISBN 978-1-4799-2495-0.

WAN, Y., Y. LIANG and L. DING. Mining Multilevel Association Rules with Dynamic Concept Hierarchy. In: *Proceedings of the Seventh International Conference on Machine Learning and Cybernetics.* Kunming, Čína: IEEE Computer Society, 2008, s. 287-292. ISBN 978-1-4244-2095-7.

Elektronické publikace

ANDERSON CONSULTING. *The Essential Guide to Day-to-Day Category Management* [online]. 2000 [cit. 4-6-2015]. Dostupné z: http://www.gs1belu.org/sites/default/files/publications/files/essential_guide_daytoday_catman.pdf

BORGELT, Christian a kol. *arules: Mining Association Rules and Frequent Itemsets* [online]. 2015 [cit. 20-3-2015]. Dostupné z: http://cran.r-project.org/web/packages/arules/index.html

GARTNER. *IT Glossary – Big data* [online]. 2013 [cit. 15-2-2015]. Dostupné z: http://www.gartner.com/it-glossary/big-data

HAHSLER, Michael and Sudheer CHELLUBOINA. *arulesViz: Visualizing Association Rules and Frequent Itemsets* [online]. 2014 [cit. 20-3-2015]. Dostupné z: http://cran.r-project.org/web/packages/arulesViz/index.html

INSIDE-R. *What is R?* [online]. 2015 [cit. 27-4-2015]. Dostupné z: http://www.inside-r.org/what-is-r

KUŽELA, Alois. *Získávání znalostí z databází* [online]. 2015 [cit. 12-12-2014]. Dostupné z: http://www.common.cz/attachments/117_alois_kuzela_asociacni_pravidla.pdfR CORE TEAM. *An introduction to R* [online]. 2015 [cit. 27-4-2015]. Dostupné z: http://cran.r-project.org/doc/manuals/R-intro.pdf

THE FACT POINT GROUP. *Leading Practices in Market Basket Analysis: How Top Retailers are Using Market Basket Analysis to Win Margin and Market Share* [online]. 2008 [cit. 11-6-2015]. Dostupné z: http://www.irgintl.com/pdf2/1.pdf

ZHAO, Yanchang. *R and Data Mining: Examples and Case Studies* [online]. 2013 [cit. 27-4-2015]. Dostupné z: http://cran.r-project.org/doc/contrib/Zhao_R_and_data_mining.pdf

Seznam zkratek

AP – Asociační pravidla

BI – Business Intelligence

CK – Cizí klíč

COFI – Co-Occurence Frequent Item

CRAN – Comprehensive R Archive Network

CSV – Comma-separated Values

DBA – Databázový administrátor

KDD – Knowledge Discovery from Databases

KPI – Key Performance Indicator

MRA – Multilevel Relationship Algorithm

OLAP – Online Analytical Processing

OLTP – Online Transaction Processing

PK – Primární klíč

POP – Point of Purchase

RFID – Radio Frequency Identification

SOFM – Self - Organizing Feature Map

TID – Identifikátor transakce

Seznam tabulek

Seznam obrázků

Prohlášení o využití výsledků diplomové práce

Prohlašuji, že

- jsem byl seznámen s tím, že na mou diplomovou práci se plně vztahuje zákon č. 121/2000 Sb. – autorský zákon, zejména § 35 – užití díla v rámci občanských a náboženských obřadů, v rámci školních představení a užití díla školního a § 60 – školní dílo;
- beru na vědomí, že Vysoká škola báňská – Technická univerzita Ostrava (dále jen VŠB-TUO) má právo nevýdělečně, ke své vnitřní potřebě, diplomovou práci užít (§ 35 odst. 3);
- souhlasím s tím, že diplomová práce bude v elektronické podobě archivována v Ústřední knihovně VŠB-TUO a jeden výtisk bude uložen u vedoucího diplomové práce. Souhlasím s tím, že bibliografické údaje o diplomové práci budou zveřejněny v informačním systému VŠB-TUO;
- bylo sjednáno, že s VŠB-TUO, v případě zájmu z její strany, uzavřu licenční smlouvu s oprávněním užít dílo v rozsahu § 12 odst. 4 autorského zákona;
- bylo sjednáno, že užít své dílo, diplomovou práci, nebo poskytnout licenci k jejímu využití mohu jen se souhlasem VŠB-TUO, která je oprávněna v takovém případě ode mne požadovat přiměřený příspěvek na úhradu nákladů, které byly VŠB-TUO na vytvoření díla vynaloženy (až do jejich skutečné výše).

V Ostravě dne ……

……

jméno a příjmení studenta